

Debt segmentation for collections and recoveries management in the Spanish market. Impact of a crisis: COVID-19 case study.

Antón Alfaya González

ABSTRACT

Collections and recoveries (C&R) activities are essential for effective credit risk management, particularly in light of post-2007-2008 financial crisis regulations. Despite their importance, this area is often underexplored. Accurate unpaid debt segmentation is crucial for optimizing collection strategies and assessing collection potential. Our objective is to identify the key determinants influencing recovery rates for unsecured debt portfolios that banks and financial institutions have written off, and to assess how these determinants were affected by the COVID-19 crisis. Specifically, we aim to identify which variables significantly predict short-term repayment behaviour before, during, and after the pandemic. To this end, we developed statistical models utilizing the CHAID classification tree methodology to identify the most discriminating variables. Our findings indicate that the COVID-19 crisis had a notable impact, causing significant alterations in these variables. The insights from this research are expected to be valuable to key stakeholders in the C&R sector, including creditors, servicers, and regulators/supervisors. In this research, we utilized a database from a prominent Investment Firm, containing information on 900.000 accounts of unsecured debt portfolios sourced from the Spanish non-performing loans market, and their payment behaviour from 2018 to 2023, along with macroeconomic data from the “Instituto Nacional de Estadística”.

Keywords: financial sector, risk management, collections and recoveries, recovery rate, debt segmentation, modelling, analytics, and COVID-19.

1. Introduction

The relevance of the credit market and C&R is significant. Credit refers to the provision of a benefit—whether in the form of a product, service, or monetary amount—with deferred payment. This practice is prevalent across sectors of the global economy, including banking, finance, retail, telecommunications, and utilities. When credit is extended for the delivery of products and services, there is inevitably a risk of non-payment, commonly referred to as credit risk. This type of risk is among the oldest in financial markets (Molina et al., 2015; Spuchlakova & Cug, 2015; Stifler & Parrish, 2014).

The primary phases of the credit risk management cycle can be articulated as three key components (Molina et al., 2015): Admission – this phase involves determining the customers eligible for the product or service, along with the specific conditions under which it will be granted, including factors such as pricing, loan amounts, instalment terms, and required guarantees; Portfolio management – following the approval of the product or service, this step entails establishing the controls and policies necessary for effective credit risk management; and Collections & recoveries (C&R) – in this phase, policies and procedures are put in place to address unpaid debts and manage situations of default.

The banking and finance sector is the most advanced and sophisticated area of credit risk management, primarily due to its inherent nature and critical significance. In comparison, sectors such as telecommunications and utilities are less developed in this regard. This distinction likely explains the greater breadth of regulations, research, and publications in banking and finance than in other

industries. Consequently, our references in this research will predominantly draw from this sector.

The financial crisis of 2008 precipitated a substantial escalation in the volume of non-performing assets, defined as financial exposures in a state of insolvency. This scenario underscored the necessity for rigorous modelling of the various facets of credit risk to facilitate sound risk management practices (Várgedő, 2024; Carleo & Rocci, 2024; Ramos Gonzalez et al., 2023; Gambetti et al., 2019).

The Basel II Accord established the framework that allows banking and financial institutions to use their internal credit risk parameters to calculate expected losses and capital requirements. This methodology is known as the internal ratings-based (IRB) approach. The three components of expected losses (EL) under the IRB approach are as follows: Probability of Default (PD) – this component indicates the percentage of borrowers expected to default within one year; Exposure at Default (EAD) – this metric estimates the amount that remains outstanding in the event of a borrower's default; and Loss Given Default (LGD) – this shows the percentage of exposure that the entity might lose if a borrower defaults.

The equation for expected losses is given by: $EL = PD * EAD * LGD$

Typically, LGD is expressed as $LGD = 1 - \text{Recovery Rate (RR)}$, where the Recovery Rate is defined as the proportion of the amount outstanding that is successfully collected or recovered by the entities, after deducting administrative fees incurred during the collection process (Galow et al., 2024; Hoechstetter & Nazemi, 2013; Repullo et al., 2010; BCBS, 2005).

Banks that employ LGD models with strong predictive capabilities can gain a significant competitive advantage, whereas models with poor predictive performance may lead to adverse selection (Gürtler & Hibbeln, 2013).

As illustrated below, the activities involved in collecting and recovering outstanding amounts are crucial to LGD, underscoring the vital role of C&R management in credit risk management.

In recent decades, increased competition among banks has influenced their credit risk, leading to a rise in non-performing loans (NPLs) (Radivojević et al., 2019). NPLs are a significant concern for banks, regulators, and market participants, as they pose risks to the overall economy and financial system. These loans negatively impact banks' profitability, consume valuable resources, and may restrict banks' ability to lend to the real economy (EBA, 2019). As a result, NPLs have garnered substantial attention from European regulators in recent years. Among their guidelines, regulators recommend that banks consolidate their NPLs and sell them to specialized investors, thereby mitigating impairment losses and addressing financial stability concerns (Bellotti et al., 2021).

Debt buyers and investors are specialized firms that acquire charged-off or delinquent debts from banks, credit card companies, and other creditors for a fraction of the total amount owed. After obtaining these debt portfolios, they strive to collect outstanding debts and, in some instances, may even sell the debts to other buyers. This practice of debt buying represents one of the most significant developments in the debt collection industry in recent years. Key components of this market include debt buyers, debt sellers, the pricing of debts—which is

influenced by factors such as demand, availability, age of the debt, frequency of prior collections or sales, geographic location, and documentation—and the agreements negotiated between buyers and sellers. These agreements define the price and face value of the debts included in the sale, the types of debts in the portfolio, the accompanying information, the accuracy of the account data, and any supporting documentation for the accounts (Stifler & Parrish, 2014).

The European market for debt collection and management services originated in the Nordic countries during the 1990s, driven largely by activities in the real estate and banking sectors. Several of the largest European debt collection agencies and investors, such as Intrum and Axactor, have their headquarters in the Nordics and maintain branches across Europe. The financial crisis of 2008-2010 spurred the growth of the servicers industry in the UK, Spain, and Italy, partially influenced by directives from European regulators. In the wake of the crisis, the NPL ratio increased notably, prompting the European Banking Authority (EBA) to issue guidelines and a Council Action Plan to reduce distressed receivables and prevent the accumulation of new distressed assets. This action plan proposed measures in three main areas: enhancing supervision; developing secondary markets, with key strategies like debt portfolio sales and securitizations to drive NPL reduction; and reforming insolvency frameworks (Malinconico & Virguti, 2024).

The onset of the COVID-19 crisis at the close of 2019, culminating in global repercussions throughout 2020, engendered a profound health, social, and economic crisis (Lin & Lu, 2024). In response to the pandemic, a range of measures was enacted to curb virus transmission and alleviate its economic and social

repercussions. These measures, which included restrictions on movement and business operations, contributed to a severe economic recession that affected nearly every sector and activity. Credit risk management, along with C&R, faced substantial challenges during this period (Lin & Lu, 2024; Ramos Gonzalez et al., 2023; O'Brien et al., 2021; Altman, 2021).

Although references to unpaid debts and C&R activities have long been prevalent in the general literature, scholarly research in this area has been relatively scarce (Malinconico & Virguti, 2024; Kriebel & Yam, 2020; Cheng & Cirillo, 2018; Thomas et al., 2016). However, interest in this subject has increased markedly in recent years (Li et al., 2023A; Sopitpongstorn et al., 2021).

Nearly two centuries ago, Balzac dedicated a novel to portraying what could be conceptualized as a manual for the "professional" debtor, illustrating methods of subsisting without repaying debts (Balzac, 1827). In stark contrast to the characters depicted in Balzac's works, C&R activities allocate considerable resources and efforts towards mitigating the repercussions of unpaid debts across various economic sectors linked to credit. In doing so, they help foster a healthier economic environment (ECB, 2017; Drozd & Serrano-Padial, 2013).

The adage "past collections do not guarantee future collections" is a well-known principle within the collections and recoveries (C&R) industry. However, analysing and comprehending past collection efforts can significantly enhance future collections by facilitating effective debt segmentation and the development of customized collections and recovery strategies tailored to each identified segment (Crespo & Govindarajan, 2018; Higginson et al., 2018).

Segmentation of debt is applied across all the different management phases (preventive, early arrears, late arrears, and written-off) for a better understanding of the debt and the debtor characteristics. Debtor profiles exhibit significant variability, encompassing categories such as prompt payers, slow payers, careless debtors, recalcitrant debtors, liar individuals, deadbeat debtors, skip debtors, and professional credit offenders, among others. A nuanced understanding of these distinct debtor profiles is imperative for categorizing delinquent accounts, assessing their collection potential, and formulating tailored strategies to optimize collections within each segment (Coleman, 2004). The details of these strategies include, among others, the type and number of amicable actions (e.g., calls, SMSs, or letters), the initiation of litigation, and the prioritization of accounts based on estimated collectability.

Higginson et al. outline seven essential pillars for best-in-class collections operations, with segmentation being the foremost. The remaining six pillars include skip tracing, contact strategy, treatment choices, frontline capabilities, performance management, and organization. The authors emphasize that the primary purpose of segmentation is to differentiate between delinquent customers who need human contact and those who can be effectively engaged through automated messaging or may require no contact at all. Furthermore, they assert that integrating machine learning and advanced analytics will enhance both value-at-risk and behavioural segmentation, thereby refining the contact and treatment strategies tailored to each defined client segment (Higginson et al., 2018).

To effectively contextualize the Spanish market, it is essential to examine the significance of the debt C&R sector by analysing data from the annual market reports published by DBK Informa, ANGEKO, and Axis Corporate. This examination will facilitate a comprehensive understanding of the dynamics and relevance of the debt C&R business within the Spanish economic landscape.

DBK Informa emphasizes the significant atomization of the Spanish market, which is characterized by the presence of over 800 companies and a notable influence from foreign firms (Informa, 2023). In its annual report for the end of 2022, similar figures to those reported by ANGEKO are noted:

- Approximately €300 million in debt under management
- €1.8 billion collected, representing 4% of the total debt
- €1.5 billion in total revenues
- Around 20,000 employees
- €14.6 billion in debt portfolios sold

ANGEKO, the National Association of Collection and Recovery Entities—“Asociación Nacional de Entidades de Gestión de Cobro”—publishes its Annual Market Study, which compiles key market figures from its members (ANGEKO, 2024). In 2023, entities affiliated with ANGEKO accounted for 85% of the debt collection and recovery sector in Spain, with the following key indicators:

- 124 million files managed
- €463,973 million in debt under management
- €12,422 million collected

- €1,748 million in total revenues, with €1,419 million derived from C&R activities
- A total workforce of 21,086 employees, of which 17,176 are involved in C&R

To establish a reference point for the sector's significance, it is noteworthy that the revenue generated by ANGEKO associates in 2023 accounted for nearly 1% of Spain's Gross Domestic Product (GDP) that year. This statistic underscores the sector's substantial contribution to Spain's overall economic landscape.

Axis Corporate analyses the Spanish market in its latest annual report (Axis Corporate, 2025). The report highlights the ongoing concentration and reconfiguration process within the C&R servicing sector, emphasizing the integration of AI across all of its activities. Furthermore, it includes a survey of key industry players, revealing that many regard AI and advanced data analysis as the most significant factors impacting their portfolio management success.

Additionally, Axis Corporate estimates that the volume of debt portfolios sold in the Spanish market will reach €17 billion in 2024, with an overall annual market size projected at approximately €15 billion.

The primary objective of this research is to deepen the understanding of debt segmentation and the influence of crises, such as the COVID-19 pandemic, on this field. Specifically, it seeks to identify the key determinants of recovery rates for unsecured debt portfolios that banks and financial institutions have written off, and to evaluate how these determinants have been affected by the COVID-19 crisis (and all the measures delivered to alleviate the pandemic impact): which

variables significantly predict short-term repayment behaviour before, during, and after the COVID-19 crisis?

To facilitate the achievement of these objectives, access to a database from a preeminent Investment Firm within the Spanish NPL market has been granted. This database comprises detailed information on unsecured debt portfolios sourced in Spain, with an average debt duration of 11 years for Individuals and 12 years for small and medium-sized enterprises (SMEs), along with pertinent account characteristics and payment behaviours documented from 2018 to 2023.

The reminder of the paper is organised as follows. Section 2 provides a summary of the literature review relevant to the research area, focusing explicitly on the C&R business and the modelling of RR and LGD to identify the main determinants of debt segmentation and the most widely used methodologies, and highlights the identified research gap. Section 3 describes the methodology employed in the research, and Section 4 presents the data used. In Section 5, research results are demonstrated. And the subsequent Section 5 presents the primary conclusions and recommendations drawn from this research.

2. Literature review

References to unpaid debts and C&R activities have been scarce in scholarly research, but interest in this topic has grown significantly in recent years (Li et al., 2023A).

As previously stated in this document, the primary phases of the credit risk management cycle include Admission, Portfolio Management, and Collections &

Recoveries (Molina et al., 2015). In practical applications of credit risk management, there has historically been a predominant focus on the admission phase, as this area is most pertinent to the business's economic viability. This trend is also reflected in the academic literature, which emphasizes the Probability of Default (PD) associated with the admission phase. At the same time, LGD, RR, and C&R have received comparatively less attention, despite being critical components of credit risk and essential for effective risk management processes. However, in recent years, there has been a noticeable increase in research concerning LGD, particularly related to the C&R phase, spurred by banking regulations introduced under Basel II and III (Várgedő, 2024; Jacobs, 2024; Loterman et al., 2012; Altman et al., 2004).

As previously mentioned in this document, C&R activities play a crucial role within the credit risk management cycle, as the cost and effectiveness of the collections process directly impact the pricing and availability of consumer credit. Despite this significance, the collections industry has garnered surprisingly little attention from economic scholars. Generally, the key aspects identified in academic literature regarding the collection of bad debt can be categorized into three main areas: locating debtors, segmentation (which involves distinguishing between consumers who cannot pay due to a lack of resources and those who will not pay even if they have the means), and the use of technology (Malinconico & Virguti, 2024; Fonseca et al., 2017; Thomas et al., 2016; Hunt, 2007). Additionally, analytics-driven intelligence is emphasized as crucial for transforming collections (Crespo & Govindarajan, 2018). The field of collections is recognized as a prime candidate for AI transformation within retail banking, given its high business impact

and significant technical feasibility (Giovine et al., 2024). Another important aspect of debt management is minimizing the cost of debt while maintaining an acceptable level of risk—striving for the lowest possible cost with a reasonable degree of prudence. Effective debt management strategies are essential for ensuring financial stability (Jonasson et al., 2019).

The earliest academic work we are aware of that examines debt collections dates back nearly a century and focuses on the debt collection market in the USA (Fedaseyeu & Hunt, 2015; Krumbein, 1924). In this study, Krumbein outlines the operations of debt collection agencies and their market dynamics during that period. This research highlights the separation of collection management from credit management following the Industrial Revolution and underscores the need to tailor collection strategies and activities to different debtor classes. This concept serves as the foundation for segmenting debts and for defining strategies to optimize C&R outcomes. Furthermore, it discusses how creditors began outsourcing debt management to third-party entities—collection agencies—and the organization and relationship that developed between these agencies and creditors. The primary objectives and requirements for collection agencies are threefold: to achieve the specified collection results, to achieve these results at minimal cost, and to foster amicable relations between creditors and debtors. Internally, collection agencies are structured by dividing their activities into two main functions: 1) client engagement and 2) debtor contact. Creditors contract these agencies based on a compensation model that may consist of a percentage of the recovered funds, a retainer fee for the duration of service, or a

combination of both. Such remuneration schemes remain prevalent today (Malinconico & Virguti, 2024; Fedaseyeu, 2015).

Several studies have analysed the impact of regulation on C&R activities and on RRs. Fedaseyeu concluded that stricter laws governing third-party debt collection led to fewer third-party collectors and a decline in RRs for delinquent credit cards (Fedaseyeu, 2020). Similarly, Fonseca et al. found that restrictions on collection practices reduce access to credit and deteriorate indicators of customers' financial health (Fonseca et al., 2017).

Within the academy, we identify three primary approaches or phases for creditors to establish their C&R strategies and processes. These options include managing debt collection internally through in-house departments or teams, outsourcing collection activities to external third-party agencies, or selling the debts. (Carleo & Rocci, 2024; Li et al., 2021; Fonseca et al., 2017; Drozd & Serrano-Padial, 2013).

C&R management is typically organized into several sequential phases, primarily delineated by the age of the delinquent debt. These phases include preventive measures, early arrears management, late arrears management, and the handling of written-off debts (Malinconico & Virguti, 2024; ECB, 2017; Bailey, 2002; Lawrence, 1992). Each phase is characterized by distinct strategies and interventions to manage the debt recovery process effectively.

Delegating the collection of NPLs to external agencies means a bank entrusts these challenging recoveries to specialized companies in exchange for a designated service fee. Third-party debt collectors often employ more aggressive collection practices than the original creditors, and their expertise generally results

in lower collection costs or higher recovery rates, as debt collection is their primary focus. Additionally, creditors typically distribute debt collection tasks among various third-party agencies, with each agency frequently representing multiple creditors (Li et al., 2022; Beck et al., 2017; Fedaseyeu & Hunt, 2015).

Debt sales typically occur after a debtor has been delinquent for 120 to 180 days, and after accounts are charged off (Athreya et al., 2018). This practice can be traced back to the 2007-2008 financial crisis, which highlighted the necessity for third-party buyers to relieve originators from the burdensome task of debt collection (Carleo & Rocci, 2024; Hoechstetter et al., 2012).

Recently, Malinconico & Virguti emphasized that there are currently five distinct business models within debt collection activities, in addition to the original creditors: a) Third-party servicing (Debt collection agencies or third-party servicers): These entities specialize in the collection and management of debt portfolios on behalf of external organizations, such as banks, financial institutions, or investment funds, without having any financial interest in the managed loans; b) Debt acquisition and recovery (Pure servicers): This model focuses on the direct purchase of debt portfolios, followed by the collection and management of these acquired assets; c) Hybrid model of debt acquisition: This approach merges elements of both debt acquisition and third-party servicing, allowing for a flexible strategy in managing debt; d) Pure investors: In this model, investors acquire debt portfolios and delegate the collection and management processes to one of the aforementioned operators; e) Special Purpose Vehicles (SPV): This category is a specific type of investor, where the acquisition of debt portfolios is facilitated

through the issuance of typically rated asset-backed securities (Malinconico & Virguti, 2024).

The primary activities involved in collections include refraining from action when payment is anticipated in the coming days, issuing a standard statement that provides a reminder, making telephone calls, sending letters, visiting borrowers in person, and pursuing legal measures. All these efforts aim to achieve one of several outcomes: recovering the full outstanding debt, collecting a portion of the debt through discount offers, establishing a payment plan, initiating legal proceedings, selling the debt, or closing the account (Thomas et al., 2012; Hoechstetter et al., 2012; Thomas et al., 2009; Rosenberg & Gleit, 1994).

Shoghi emphasizes the significance of employing AI and machine learning technologies in the collection and analysis of large datasets. He argues that these advanced technologies are crucial for effectively classifying debtors and systematically recommending appropriate actions regularly (Shoghi, 2019).

In the context of research on LGD and RR, two primary streams have been identified. The first examines the relationship between Probability of Default (PD) and LGD-RR, while the second focuses on estimation methodologies and the key determinants influencing both LGD and RR (Han & Jang, 2013; Qi & Zhao, 2011; Bastos, 2010).

Altman et al. analysed the evolution of literature regarding the relationship between Probability of Default (PD) and LGD-RR. Their findings indicate that earlier research suggested an inverse relationship between PD and RR. Subsequent studies presented the view that RR were fixed-ratio, rendering PD and RR

independent. More recent research, however, demonstrates that PD and RR are negatively correlated (Altman et al., 2005; Altman et al., 2004).

Subsequent research in this area indicates a correlation between default rates, LGD, and RRs. Specifically, a negative correlation is observed for RRs, whereas a positive correlation is observed with LGD (Calabrese, 2012; Bellotti & Crook, 2012). However, some studies report differing results that vary with the stage of the credit cycle, whether during upturns or downturns (Jimenez & Saurina, 2006).

Furthermore, some researchers have expanded their inquiries into the relationship between the probability of default (PD) and LGD-RR while accounting for the macroeconomic environment. During periods of recession or industry downturns, default rates tend to rise while RRs decline. Although there is a stronger correlation between default rates and RRs than with macroeconomic variables, this relationship reveals common dynamic patterns that are associated with the business cycle (Várgedő, 2024; Li et al., 2023A; Bruche & González-Aguado, 2010).

Regarding determinants and methodology, there is no consensus on which are most suitable for modelling LGD and RR. However, various studies shed light on relevant factors and estimation methods (Han & Jang, 2013).

Estimating and forecasting the LGD and RR values present significant challenges (Wu et al., 2024; Jacobs, 2024; Kellner et al., 2022; Thomas et al., 2012). The literature identifies several reasons for these difficulties: the complex circumstances of individual borrowers (Li et al., 2023A); the variety of influencing variables, the prolonged duration of the recovery process, and the diverse nature of the distributions involved (Li et al., 2022); limited data availability, particularly

from third-party purchasers, repayment rates typically exhibit a bimodal distribution, with high probabilities clustered near 1 and 0, while collection workout costs remain undisclosed (Nazemi et al., 2022; Bastos, 2010; Grunert & Weber, 2005); the distributions of values often display high irregularity (Li et al., 2021); the RR itself is a continuous variable that is fractional and bounded within $[0,1]$, characterized by bimodal and asymmetric density, necessitating value transformation for statistical purposes (Sopitpongstorn et al., 2021); and finally, effective modelling requires additional data, the application of more advanced techniques, consideration of a greater number of influencing factors, and unfortunately, the current data availability is limited (Chalupka & Kopecsni, 2008).

A consensus on the most effective methodology has yet to be established (Li et al., 2023A). Historically, various regression techniques have been employed, with ordinary least squares (OLS) regression serving as the benchmark (Li et al., 2023A; Ye & Bellotti, 2019; Bellotti & Crook, 2012).

The primary methodological approaches include various linear and non-linear regression methods, applied at one, two, or multiple stages. However, no consensus has emerged regarding which method performs best, as findings differ across various academic studies. A selection of commonly used methodologies from the literature is presented in Appendix B.

Decision and regression trees have emerged as compelling alternatives to the parametric models typically used in empirical studies of LGD and RR, such as logistic regression. In some research, these tree-based methodologies have even

demonstrated superior performance compared to other approaches (Bastos, 2014; Bastos, 2010; Matuszyk et al., 2010; Rosenberg & Gleit, 1994).

Decision trees were first developed in the early 1960s by H. Raiffa and his colleagues at Harvard Business School. Subsequently, Sparks, Breiman, and Makowski advanced these concepts in the years that followed, particularly applying them to credit scoring. One effective way to utilize a decision tree is to associate each node with the probability of non-payment for the individuals represented by that node. The probability at a given node, denoted as node n , is determined as a weighted average of the probabilities from its child nodes, where the weights correspond to the number of individuals at node n . To decide an observation (or account), we traverse the tree from the root node, selecting the appropriate paths based on the observation until we reach the relevant leaf node. By comparing the probability of non-payment at this node to a predetermined cutoff, we can arrive at a decision. However, a significant drawback of decision trees is the larger sample sizes required to obtain statistically reliable probability estimates at each node. Nevertheless, they offer substantial advantages, such as effectively reflecting the impact of various combinations of factors, allowing for cost-structure variability across nodes, and providing overall ease of use and understanding (Bastos, 2014; Rosenberg & Gleit, 1994).

Machine learning (ML) is increasingly being applied to forecasting challenges in economics and finance, emerging as a more effective and practical approach. Financial institutions, including banks and fintech companies, have been incorporating ML techniques into their operations (Li et al., 2024; Nazemi & Fabozzi,

2024; Dumitrescu et al., 2022; Kellner et al., 2022). This trend is also reflected in the literature on LGD and RR, where recent studies have compared traditional regression methodologies with newer ML techniques such as XGBoost, neural networks (NN), and random forests (RF).

In a recent study, it was found that classification trees and random forests achieved the highest accuracy compared with machine learning tree-based algorithms (such as decision trees, random forests, light gradient boosting, and XGBoost) and neural network algorithms (including backpropagation neural networks and extreme learning machines). Notably, XGBoost and extreme learning machine techniques exhibited the best performance (Li et al., 2024).

Regarding the determinants of LGD and RR, there is no consensus in the literature on the most significant factors. Additionally, studies in this field often yield contradictory conclusions (Xu et al., 2024; Radivojević et al., 2019).

Returning to the academic literature on credit scoring, we identify five primary information sources integral to scoring systems, commonly referred to as the 5 Cs of credit: Character (the borrower's willingness to repay their debt), Capacity (the financial ability of the borrower to meet repayment obligations), Capital and collateral (the assets or equities of the borrower that can be liquidated for payment), and Conditions (which pertain to the overall economic environment or specific circumstances affecting the borrower or type of credit) (Rosenberg & Gleit, 1994).

In a similar vein to the initial 5 Cs of credit, our review of the literature on the determinants of LGD and RR reveals four key sources of information regarding

these determinants: borrower characteristics, which encompass the principal traits of the debtor at the time credit was extended and how these have evolved, such as age, educational background, and sector of activity; debt or account characteristics, which involve all variables related to the debt, account, or contract, including debt amount, product type, and the presence of collateral; macroeconomic variables, such as GDP, unemployment rate, and interest rates; and repayment behaviour and C&R activities, which cover the actions taken by creditors to recover unpaid debts, including trace/no trace, litigation, and the number of late payments. We have compiled references related to these four primary sources in Table 1.

Table 1. Summary of data sources references for LGD and RR determinants

Data sources	Authors
Borrower characteristics	Xu et al., 2024; Li et al., 2024; Li et al., 2023A; Li et al., 2023B; Nazemi et al., 2022; Li et al., 2022; Calabrese & Zanin, 2022; Bellotti et al., 2021; Wang et al., 2020; Ye & Bellotti, 2019; Shoghi, 2019; Serrano-Cinca et al., 2015; Bijak & Thomas, 2015; Han & Jang, 2013; Bellotti & Crook, 2012; Hoechstetter et al., 2012; Khieu et al., 2012; Höcht & Zagst, 2010; Chen et al., 2009; Chalupka & Kopecsni, 2008; Grunert & Weber, 2005
Debt or account characteristics	Xu et al., 2024; Li et al., 2024; Li et al., 2023A; Li et al., 2023B; Nazemi et al., 2022; Li et al., 2022; Calabrese & Zanin, 2022; Bellotti et al., 2021; Wang et al., 2020; Ye & Bellotti, 2019; Shoghi, 2019; Serrano-Cinca et al., 2015; Bijak & Thomas, 2015; Han & Jang, 2013; Bellotti & Crook, 2012; Hoechstetter et al., 2012; Khieu et al., 2012; Höcht & Zagst, 2010; Chen et al., 2009; Chalupka & Kopecsni, 2008; Grunert & Weber, 2005
Macroeconomic variables	Xu et al., 2024; Li et al., 2023A; Li et al., 2023B; Nazemi et al., 2022; Li et al., 2022; Bellotti et al., 2021; Wang et al., 2020; Bijak & Thomas, 2015; Han & Jang, 2013; Bellotti & Crook, 2012; Khieu et al., 2012; Höcht & Zagst, 2010; Grunert & Weber, 2005
Repayment behaviour and C&R activities	Xu et al., 2024; Li et al., 2024; Li et al., 2023A; Li et al., 2023B; Nazemi et al., 2022; Li et al., 2022; Calabrese & Zanin, 2022; Bellotti et al., 2021; Wang et al., 2020; Ye & Bellotti, 2019;

	Shoghi, 2019; Serrano-Cinca et al., 2015; Bijak & Thomas, 2015; Han & Jang, 2013; Bellotti & Crook, 2012; Hoechstetter et al., 2012; Khieu et al., 2012; Höcht & Zagst, 2010; Chen et al., 2009; Grunert & Weber, 2005
--	--

Source: Own elaboration

Basel II recommends that banks account for macroeconomic downturn conditions when predicting LGD and RR. Numerous studies have examined the correlation between these factors, yielding mixed results and no consensus. However, most findings suggest that incorporating macroeconomic variables enhances predictive accuracy.

Utilizing these macroeconomic variables helps capture the impact of business cycles on loan repayment and collection processes (Xia et al., 2021). Furthermore, economic uncertainty is considered a critical systematic determinant of RR distributions (Gambetti et al., 2019). Some studies have even concluded that the determinants of RR differ between "good" and "bad" times within the credit cycle, indicating that loan characteristics play a more significant role during prosperous periods, while specific firm characteristics and collateral determinants become more influential in challenging times (Wang et al., 2020).

The macroeconomic factors most frequently examined in studies include the unemployment rate, which typically exhibits a negative relationship with RR (Galow et al., 2024; Xia et al., 2021; Calabrese, 2012). In contrast, gross domestic product (GDP) generally demonstrates a positive relationship with RR (Galow et al., 2024; Radivojević et al., 2019; Friedman & Sandow, 2005). Interest rates usually have a negative correlation with RR (Galow et al., 2024; Xia et al., 2021). Stock market returns are generally positively related to RR (Galow et al., 2024; Bruche &

González-Aguado, 2010), while investment growth also tends to show a positive association with RR (Bruche & González-Aguado, 2010). Household final consumption expenditure typically shows a negative relationship with RR (Radivojević et al., 2019). In contrast, the consumer confidence index shows mixed results, with either a negative or a positive correlation with RR (Xia et al., 2021). Appendix C summarizes the primary literature on the inclusion of these macroeconomic variables and their impact on model predictive power.

The characteristics of borrowers and their debt or account attributes constitute broader categories of determinants that influence LGD and RR and carry significant weight in the statistical models used to predict them. This category encompasses various factors including whether the borrower is a natural or legal person, the number of days in arrears, instances of restructured loans, the debtor's age and income, residential status, marital status, educational background, address details, duration of relationship with the bank, type of financial product, years residing at the current address, employment status, credit bureau score, agency ratings, age of the defaulted account, loan amount, loan term, and the presence of multiple loans and/or creditors. Additionally, it considers attributes such as firm size, industry, age, legal structure, financial ratios, collateral or personal guarantees, type of collateral, debtor address region, borrower's indebtedness, EAD, Loan-to-Value (LTV) ratio, overall risk level, purpose of the loan, currency type, loan interest rate, customer outlook, country of residence or incorporation, instances of financial hardship, borrower affordability, and borrower viability. A summary of the most pertinent references found in the relevant literature regarding these variables is presented in Appendix D.

In recent years, there has been a growing emphasis on repayment behaviour and actions taken in the C&R process, particularly on their contributions to research on RR and LGD. Information related to the C&R process has been shown to enhance the performance of LGD and RR models significantly. This type of information encompasses various data points, including the duration of centralized management, the involvement of external agencies, the initiation of litigation, whether or not there is traceability of debtors, the duration and costs associated with the workout process, credit bureau information, communication records (calls and contacts), the reasons for default, and repayment behaviour statistics (such as the number of late payments, average late payment duration, payment discrepancies, and cumulative differences in payments). Notably, researchers have placed special emphasis on the importance of post-default information for further enhancing the predictive capabilities of these models. Appendix E summarizes the primary references to these variables in the existing literature.

While the existing literature provides valuable insights into C&R activities, methodologies, and the determinants of RR and LGD, we have not identified any studies that directly address the specific focus of our research.

To the best of our knowledge, this study represents the first exploration of the most effective segmentation variables and the impact of an event such as COVID-19 (and all the linked measures to mitigate its effect) on debt segmentation for collections and recoveries of unsecured written-off portfolios acquired by investment firms from banks and other financial institutions in the Spanish market.

This study contributes to the research area of the more relevant determinants for RR modelling estimations, as our models predict the probability of an account to make a payment in the following 30 days, directly linked to the RR.

The results of our research align with recent developments in the literature regarding the relevance of the four sources of information (borrower characteristics, debt characteristics, macroeconomic variables, and repayment behaviour and C&R activities), as well as the increasing importance of the C&R activities variables. But our research goes further, emphasizing the relevance of this category of variables.

On the contrary, our results are completely distinct from the more traditional literature, which only notes the importance of borrower and debt characteristics and does not include the C&R activities variables.

It is relevant to highlight that our research uses data that is typically inaccessible to academic studies because it is not publicly available. The dataset is comprehensive, encompassing information from approximately 900,000 accounts and tracking their payment behaviour over six years, from 2018 to 2023. This database draws upon the four primary sources identified in the literature: borrower characteristics, debt characteristics, macroeconomic variables, and repayment behaviour and C&R activities.

3. Research methodology

3.1 Research framework

This research was developed in five phases, beginning with a literature review of C&R, particularly focusing on RR and LGD modelling. The aim was to identify key debt segmentation determinants and assess effective modelling methodologies. The review aligned academic findings with practitioners' approaches, resulting in a comprehensive list of determinants and methodologies that integrated insights from both sources.

Next, data were gathered from two sources: the Investment Firm systems, for account and debtor characteristics, payment behaviour (2018-2023), and collection activities; and the INE public website for macroeconomic factors. Data cleaning and normalization were conducted to enhance analysis efficiency.

In the third phase, we developed statistical models by initially creating the dependent variable: the probability that an account makes a payment within 30 days. Additionally, we derived additional variables and performed univariate and bivariate analyses to assess their correlation with the dependent variable. The dataset was divided into Individuals and SMEs, using 80% for training and 20% for testing.

The models were developed, using CHAID classification trees, to differentiate between Individual and SME debtors, acknowledging their distinct characteristics and payment behaviours. To analyse the variation in the importance of determinants influenced by the impact of COVID-19, several models were created: one model encompasses the entire period from 2018 to 2023 ("Global"); a second model covers the timeframe preceding the pandemic, specifically 2018-2019 ("Pre-COVID"); the third model addresses the period during the

pandemic, which includes 2020-2021 (“COVID”); and, lastly, a model is dedicated to the timeframe following the pandemic, covering 2022-2023 (“Post-COVID”). In total, eight models were constructed, as detailed in Table 2.

Table 2. Summary of constructed models

Segment	Models
Individuals	• Individuals Global model • Individuals Pre-COVID model • Individuals COVID model • Individuals Post-COVID model
SMEs	• SMEs Global model • SMEs Pre-COVID model • SMEs COVID model • SMEs Post-COVID model

Source: Own elaboration

Once the models were developed, we analysed the variables included in each model and assessed their importance weight. The final number of variables ranged from 19 to 26, as follows: Individuals Global – 25 variables; Individuals Pre-COVID – 26; Individuals COVID – 24; Individuals Post-COVID – 22; SMEs Global – 21; SMEs Pre-COVID – 19; SMEs COVID – 24; and SMEs Post-COVID – 19. However, the significance of these variables was primarily concentrated on 9 to 12 of them, with the remaining variables offering marginal value in enhancing the models’ predictive capabilities.

The models were initially developed with all variables, including clients with prior payments. Due to high correlation with these payments, they were revised to exclude payers. A streamlined version was created with only the most relevant variables, and all models were analysed for predictive quality using key performance indicators like the Gini coefficient and Kolmogorov-Smirnov test.

In the fourth phase, we shared the models with practitioners to gather feedback and compare results with their experiences, leading to adjustments for improvement.

Finally, we analysed and compared models to assess whether the main determinants for segmenting debt portfolios were consistent across Global, Pre-COVID-19, COVID-19, and Post-COVID-19 models, and drew conclusions.

3.2 Modelling methodology

As delineated in the Literature Review section of this document, there exists a notable lack of consensus within the academic community regarding the most effective methodology for estimating probabilities in data collection. Nevertheless, the predominant trends in methodological approaches can be categorized into three principal types: traditional regression techniques, classification and regression trees (CART), and advanced machine learning methodologies, which encompass deep learning, random forests, and boosting algorithms.

The foundational work of Breiman and Friedman in 1973 introduced recursive partitioning algorithms, with classification and regression trees (CART) emerging as the most frequently employed tree algorithms, demonstrating robust performance across various applications (Malikidou & Strohbach, 2025; Xu et al., 2024; Thomas & Bijak, 2015; Bijak, 2014). Notably, classification trees, in conjunction with random forests, have frequently achieved superior accuracy compared with alternative methods (Malikidou & Strohbach, 2025; Li et al., 2024).

Regression trees represent a potent yet straightforward technique for regression analysis, wherein the predicted values of the target variable are derived through a series of sequential logical if-then conditions. This methodical sequence of binary splits effectively partitions observations based on loan characteristics. The algorithm initiates with a root node encompassing all observations, subsequently searching through all binary splits of available attributes to identify the split that minimizes the intra-subset variation of the target variable in the resulting daughter node. Consequently, each daughter node demonstrates a greater homogeneity of the target variable than its parent node. This recursive procedure continues until no further reduction in the variation of the target variable is attainable (Bastos, 2014; Bastos, 2010).

For the present research, the chosen methodological framework is a CHAID (Chi-square Automatic Interaction Detector) classification tree. This methodology employs a decision tree algorithm primarily intended for classification and segmentation tasks, and it is notably characterized by its approach to tree construction, which accommodates non-binary divisions. Nodes in the tree are divided into two or more branches, in contrast to binary methods. Within this framework, the dependent variable must be categorical (either nominal or ordinal), while the predictor independent variables with fewer cases are combined to ensure that the resulting daughter nodes meet a minimum required size.

For each node, CHAID assesses every available predictor variable. It starts by merging the categories of each predictor that do not significantly differ

concerning the dependent variable. Then, it evaluates all possible splits of the predictor (with or without merged categories) and uses the Chi-square test or the likelihood ratio to identify the predictor that has the most significant relationship with the dependent variable at that particular node. The variable exhibiting the highest Chi-square value, along with a p-value below a predetermined threshold (defaulting to 0.05), is chosen to split the node. This process is repeated until no predictors are deemed significant or until a stopping criterion, such as minimum node size, is met.

This approach is distinguished by its robustness in the presence of outlier data, a common occurrence in collection activities. Additionally, the stability of the tree structure is maintained even as individual variables within the models evolve over time. The clarity and interpretability of these models, due to their transparency regarding determinants and their respective weights and importance, render them accessible. Furthermore, the implementation and maintenance of CHAID models are comparatively straightforward, especially when juxtaposed with more complex methodologies such as XGBoost or random forests. Beyond its academic recognition, CHAID is widely employed in business practices, particularly within the Analytics Departments of institutions engaged in collections and recovery activities, including banks, investment funds, and debt collection agencies.

The general specifications for the model's construction utilized in this research are detailed in Appendix F.

In line with emerging trends in literature regarding the application of machine learning techniques, we explored various algorithmic methodologies for this

research, including Random Forest and XGBoost (Malikidou & Strohbach, 2025; Xu et al., 2024). We evaluated these methodologies against the CHAID classification tree, focusing on their predictive power and ease of implementation, ultimately deciding to exclude them from further consideration. The primary distinction of the CHAID tree, in comparison to other ensemble methods such as Random Forest or Boosting, lies in its capacity to elucidate interactions among multiple predictors in a transparent and visually interpretable manner. This characteristic facilitates a deeper understanding of the relationships within the data, allowing for the identification of complex interactions that may not be readily apparent in more opaque models.

The tests indicated that the Random Forest model did not yield any improvements over the CHAID tree. This may be attributed to several factors, such as the CHAID tree already being effectively optimized and pruned, resulting in a favourable balance between bias and variance. Alternatively, it is possible that the hyperparameters of the Random Forest did not enable sufficient decorrelation among the various trees.

In the comparison of CHAID and XGBoost, it was noted that the Boosting model demonstrated superior performance indicators. However, the enhancement was minimal, with less than a one-point increase in the Gini coefficient. Consequently, the decision was made to proceed with the CHAID tree for several reasons: its simplicity in model development, time efficiencies in execution, ease of implementation in a production environment, straightforward maintenance, and clarity in analysing results and identifying the most relevant variables.

4. Data

For this research, we were granted access to data on portfolios acquired in the Spanish market by the Investment Firm and its affiliated entities (hereinafter, the "Investment Firm"). This Investment Firm is part of a Global Investment Firm Group that maintains a prominent presence and reputation in the market.

The research incorporates data from approximately 900,000 accounts across 35 portfolios acquired by the Investment Firm in the Spanish market between 2015 and mid-2022. These accounts are categorized into two primary groups due to their distinct behaviours (Dumitrescu et al., 2022; ECB, 2017): Individual account holders, which comprise over 700,000 accounts and represent €3.3 billion in debt; and Small and medium-sized enterprises (SMEs), which account for more than 200,000 files and €10.9 billion in debt.

The database assembled for this research includes five key sources of information, akin to those found in the academic literature (Xu et al., 2024; Li et al., 2024; Li et al., 2023A; Khieu et al., 2012). The Investment Firm provided data on accounts, debtors, payment behaviours, and collection activities. At the same time, macroeconomic factors were sourced from the public information website of the Instituto Nacional de Estadística (INE, Spanish National Statistics Institute, www.ine.es). Detailed information about these data sources is provided in Table 3, and Appendix G comprises a detailed description of the data fields.

Table 3. Data fields included in the database

Data category	Data fields
Accounts characteristics	• Account ID unique number • Individuals or SMEs segment flag

at the portfolios acquisition	• Name of the acquired portfolio • Name of the credit originator • Debt type of product • Detail of the original debt: total amount, principal amount, interests amount and others amount • Detail of the actual debt: total amount, principal amount, interests amount and others amount • Debt delinquency date • Debt cancellation date • Debt upload by the investor date
Debtor characteristics	• Debtor ID unique number • Type of participant in the debt • Date of birth • Type of ID document • Number of mobile phone numbers • Number of fixed-line phone numbers • Postal code
Payment behaviour in the 2018-2023 period of time	• Payment unique ID number • Payment amount • Payment date • Method of payment • Management phase in which the payment was made
Collection activities delivered by the Investment Firm in order to recover the debt	• SMS sending unique ID number • SMS sending date • Communications model • Calls unique ID number • Number of calls by month • Number of useful contacts by month • Number of payments promises by month • Judicial information: Judicial procedure unique ID number, Judicial procedure status, Judicial claim date, Judicial "auto" number and Judicial procedure discharge date
Macroeconomic factors	• General Consumer Price Index (on annual variations and on monthly variations) • Underlying Consumer Price Index (on annual variations and on monthly variations) • Consumer Confidence Index (ICC, actual situation and expectations) • GDP Market Prices (annual and quarterly variations) • GDP Full-Time Equivalent Employment (annual and quarterly variations) • Unemployment rate (on annual variation) • Activity rate (on annual variation)

Source: Own elaboration

5. Results

5.1- Results Summary and Confirmation of the Research Question

The models developed demonstrated strong predictive capabilities, with Gini coefficients ranging from 0,522 to 0,713 and K-S indicators from 0,379 to 0,577, compared to literature values of 0,7076 for Gini and 0,5647 for K-S (Dumitrescu et al., 2022).

The tables summarize our model variables and their statistical significance, illustrated with a heat map. We also categorized the weights into four types: C&R activities, debt characteristics, debtor characteristics, and macroeconomic factors.

Table 4. Summary of variables weight and “heat” map of the Individuals models

Global		Pre-COVID		COVID		Post-COVID			
Variable	Weight	Variable	Weight	Variable	Weight	Variable	Weight		
Judicial flag	0,5570	Judicial flag	0,4752	Creditor grouped	0,3230	Creditor grouped	0,5901		
RPC I90d	0,2240	Initial other debt amount	0,2476	Judicial flag	0,3202	Judicial flag	0,1824		
Creditor grouped	0,1336	Creditor grouped	0,1709	RPC I90d	0,3143	RPC I90d	0,1636		
Initial other debt amount	0,0693	National nationality flag	0,0514	Monthly debt amount grouped	0,0239	Product type grouped	0,0201		
GDP FTEE annual deciles	0,0116	Youngest DOB	0,0314	Calls ever flag	0,0097	Other nationality flag	0,0192		
National nationality flag	0,0034	Region	0,0122	RPC ever flag	0,0050	Calls I90d flag	0,0140		
GCPI annual deciles	0,0011	GCPI monthly deciles	0,0075	UCPI monthly deciles	0,0039	UCPI annual	0,0054		
GCPI monthly deciles	0,0000	Initial interest debt amount	0,0039			RPC ever flag	0,0032		
AR annual deciles	0,0000					GCPI annual	0,0020		
						Calls ever flag	0,0000		

Global		Pre-COVID		COVID		Post-COVID	
Type of variable	Weight	Type of variable	Weight	Type of variable	Weight	Type of variable	Weight
C&R activities	0,7810	C&R activities	0,4752	C&R activities	0,6492	C&R activities	0,3632
Debt characteristics	0,2029	Debt characteristics	0,4224	Debt characteristics	0,3469	Debt characteristics	0,6102
Debtor characteristics	0,0034	Debtor characteristics	0,0950	Debtor characteristics	0,0000	Debtor characteristics	0,0192
Macroeconomic	0,0127	Macroeconomic	0,0075	Macroeconomic	0,0039	Macroeconomic	0,0074

Source: Own elaboration

Table 5. Summary of variables weight and “heat” map of the SMEs models

Global		Pre-COVID		COVID		Post-COVID	
Variable	Weight	Variable	Weight	Variable	Weight	Variable	Weight
Guarantor flag	0,5130	Guarantor flag	0,4256	Guarantor flag	0,4605	Guarantor flag	0,4763
Judicial flag	0,3532	Judicial flag	0,3037	Judicial flag	0,3594	Judicial flag	0,2249
Product type grouped	0,0457	Monthly debt amount grouped	0,0895	Initial principal debt amount	0,1068	RPC ever flag	0,2000
GDP MP annual	0,0233	Product type grouped	0,0505	Creditor grouped	0,0187	UCPI annual	0,0252
Creditor grouped	0,0184	# phone numbers	0,0446	CCI expectations deciles	0,0160	Product type grouped	0,0236
Monthly debt amount grouped	0,0178	CCI expectations deciles	0,0220	CCI	0,0137	GCPI monthly	0,0189
# mobile phone numbers	0,0140	Calls ever flag	0,0204	GDP MP quarter deciles	0,0082	Creditor grouped	0,0166
Initial interest debt amount	0,0112	Region	0,0201	CCI actual deciles	0,0078	# fixed line phone numbers	0,0095
# phone numbers	0,0032	GCPI monthly	0,0134	GDP MP annual deciles	0,0046	GCPI annual	0,0050
		Creditor grouped	0,0088	Initial interest debt amount	0,0044	Monthly debt amount grouped	0,0000
		CCI actual	0,0014	UR annual	0,0000		

Global		Pre-COVID		COVID		Post-COVID	
Type of variable	Weight	Type of variable	Weight	Type of variable	Weight	Type of variable	Weight
C&R activities	0,3704	C&R activities	0,3687	C&R activities	0,3594	C&R activities	0,4344
Debt characteristics	0,6061	Debt characteristics	0,5744	Debt characteristics	0,5904	Debt characteristics	0,5165
Debtor characteristics	0,0000	Debtor characteristics	0,0201	Debtor characteristics	0,0000	Debtor characteristics	0,0000
Macroeconomic	0,0233	Macroeconomic	0,0368	Macroeconomic	0,0503	Macroeconomic	0,0491

Source: Own elaboration

By analysing the available data and developed models, we identify the discriminant variables affecting recovery rates. Our findings confirm that these variables have been influenced by COVID-19 in both Individual and SME segments. The pandemic clearly altered the significance of these variables, with a more pronounced effect in the Individuals models.

5.2- General Results of the Research

In analysing Individuals models, notable distinctions appear across time frames. In the Pre-COVID model, the Judicial flag is the most significant variable, followed by Initial other debt amount and Creditor. Variables related to debtor characteristics, such as National nationality flag, Youngest date of birth, and Region, along with

General-Consumer-Price-Index and Initial interest debt amount, are included, though their contributions are marginal.

In contrast, the COVID Individuals model illustrates a shift in variable importance, highlighting three key variables: Creditor, Judicial flag, and Right-Party-Contact last 90 days, which carry similar weights. The Monthly debt amount follows as significant, while Calls ever flag, Right-Party-Contact ever flag, and Underlying-Consumer-Price-Index have marginal importance. This model emphasizes collections-related variables and shows reduced focus on debtor characteristics compared to the Pre-COVID model.

In the Post-COVID Individuals models, the key variables—Creditor, Judicial flag, and Right-Party-Contact last 90 days—remain significant but have shifted in weight. Other variables include Product type, Other nationality flag, and Calls last 90 days; while Underlying-Consumer-Price-Index, Right-Party-Contact ever flag, General-Consumer-Price-Index, and Calls ever flag show marginal significance.

Within SMEs models, the Guarantor flag and the Judicial flag are key variables, though their weights vary across models. Moreover, the four variable categories—debt characteristics, debtor characteristics, collection activities, and macroeconomic factors—vary in significance.

In the Pre-COVID SMEs model, key variables include the Guarantor flag, Judicial flag, and Monthly debt amount. Additional variables are Product type, Number of phone numbers, Consumer-Confidence-Index expectations, Calls ever flag, Region, and General-Consumer-Price-Index, with Creditor and Consumer-Confidence-Index actual showing marginal significance.

Transitioning to the SMEs COVID model reveals a significant shift in variable importance. The Guarantor flag, Judicial flag, and Initial principal debt amount are the top three variables, with greater weight than in the Pre-COVID model. Subsequent variables, like Creditor and Consumer-Confidence-Index expectations, receive lower weights. Notably, the weight of collection activities has declined compared to the Pre-COVID model.

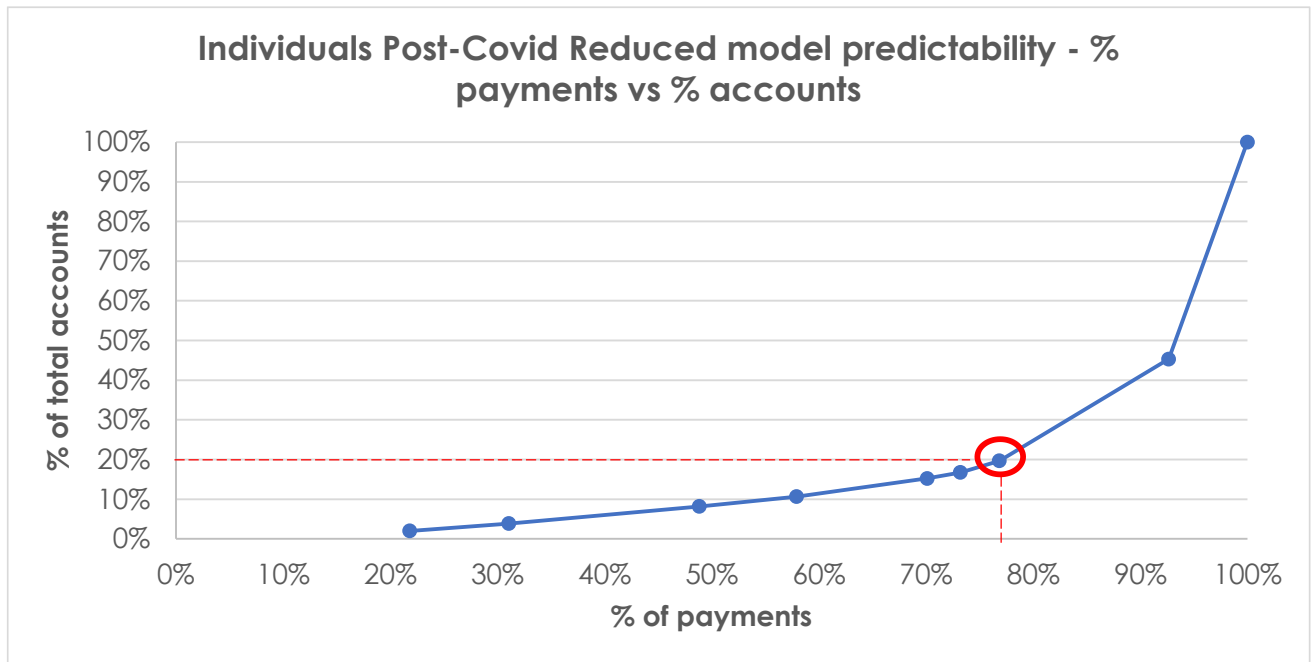
Lastly, the SMEs Post-COVID model emphasizes the Guarantor flag, Judicial flag, and Right-Party-Contact ever flag as primary variables, alongside factors like Underlying-Consumer-Price-Index annual, Product type, monthly General-Consumer-Price-Index, and Creditor. It also considers variables such as the Number of fixed line numbers, annual General-Consumer-Price-Index, and Monthly debt amount, though they hold marginal weight.

5.3- Illustrative Application of the Models Built

Segmentation models in C&R are crucial for resource optimisation. For illustration, our Individuals Post-COVID model, with a 20% sample, predicts 77% of the total accounts that will pay in the following 30 days, highlighting the model's effectiveness in improving recovery strategies.

Table 6 and Figure 1. Distribution of accounts and payments on the Individuals Post-COVID model

MODEL SCORE	% PAYMENTS	% ACCOUNTS
0,000	100%	100%
0,001	93%	45%
0,002	77%	20%
0,003	73%	17%
0,004	70%	15%
0,005	58%	11%
0,006	49%	8%
0,007	31%	4%
0,016	22%	2%



Source: Own elaboration

6. Conclusions

6.1- Discussion of the Research Results

The primary sources for LGD-RR modelling include borrower characteristics, debt characteristics, macroeconomic variables, and C&R activities. Recent literature (Xu et al., 2024) and our research confirm their relevance for both Individuals and SMEs segments, though highlighting notable differences between the segments and earlier studies.

In the literature, borrower and debt characteristics are frequently discussed, macroeconomic factors occasionally, and some studies recently include repayment behaviour and C&R variables (Xu et al., 2024; Li et al., 2023A). Our models emphasize debt characteristics and C&R variables for both segments, diverging from the traditional focus on borrower and debt characteristics.

Borrower characteristics are minimal features in our models. In the SMEs model, only Region appears in the Pre-COVID version. The Individuals model includes Nationality, Age, and Region in the Pre-COVID version, but these characteristics are absent in the COVID model, and have limited presence in the Post-COVID and Global models.

Integrating C&R activities has improved our modelling, aligning findings from previous authors (Bellotti et al., 2021; Han & Jang, 2013), which show that including debt collection activities in LGD modelling enhances predictive accuracy over models using only firm-specific variables.

Our conclusions differ from much of the existing literature mainly due to our focus on written-off portfolios with an average debt age of 11-12 years, compared to the younger debt profiles typically examined in other studies, and the limited access to information on C&R activities.

In RR modelling, borrower characteristics are more significant for newer debts, while C&R activities gain importance for older debts, reducing the impact of borrower traits. Access to repayment behaviour and C&R data is typically limited, but having this information significantly enhanced our modelling efforts.

In C&R practices, incorporating macroeconomic variables into RR modelling is less common, yet our research shows they significantly enhance predictive power, consistently appearing in our final models. However, as seen in previous literature (Han & Jang, 2013), their impact is reduced.

Our models incorporate key macroeconomic variables from established academic references, including the Consumer-Price-Index, GDP, Consumer-Confidence-Index, and activity/unemployment rates (Xu et al., 2024; Galow et al., 2024; Xia et al., 2021).

6.2- Recommendations Based on the Research Results

This research findings are valuable academically and for key stakeholders in the C&R sector: creditors, servicers, regulators, and supervisors.

For creditors and servicers managing debt portfolios, modelling potential RR enhances efficiency and effectiveness in portfolio management and valuation. It allows targeted collection efforts on accounts with higher recovery likelihood, reducing wasted resources on less collectible accounts. Additionally, it improves understanding of potential recoveries, enhancing pricing accuracy for portfolio transactions.

Incorporating debt characteristics, debtor traits, C&R activities, and macroeconomic factors in modelling is essential, and remarkably, in our research, debt characteristics and C&R activities are the key sources. Focus on variables like the Judicial flag, Creditor, and Right-Party-Contact for Individual accounts, and the Judicial flag, Guarantor flag, and Right-Party-Contact for SMEs. Additionally, entities should enhance their RR segmentation models, especially in response to significant events like the COVID-19 pandemic, which significantly affect Individuals accounts more than SMEs.

This research emphasizes the significance of segmentation models in C&R for regulators and supervisors. These models enhance C&R efficacy and efficiency by considering the key sources highlighted, ultimately supporting a stronger economy. The COVID-19 pandemic has influenced RR determinants, indicating that best practices in segmentation must adapt to future challenges to strengthen creditor resilience during downturns.

APPENDICES

Appendix A.- Glossary of terms

AdaBoost – Adaptive boosting

AI – artificial intelligence

ANN – artificial neural networks

AR – activity rate

BCBS – Basel Committee on Banking Supervision

BC-OLS – Box-Cox transformation/OLS

BNB – Bernoulli naive bayes

B-OLS – beta transformation/OLS

BR – beta regression

CART – classification and regression tree

C&R – collections and recoveries

CCI – Consumer Confidence Index

CHAID – Chi-square automatic interaction detector

CRPH – Competing risks proportional hazards

DNN – deep neural network

DT – decision tree

EAD – exposure at default

EBA – European Banking Authority

ESCB – European System of Central Banks

EL – expected losses

GBDT – gradient boosting decision trees

GDP – Gross Domestic Product

GDP-FTEE – GDP Full-Time Equivalent Employment

GDP-MP – GDP Market Price

GEV – generalised extreme value

GLM – generalized linear model

GNB – Gaussian naive bayes

INE – “Instituto Nacional de Estadística”, National Institute of Statistics

KNN – K-nearest neighbours

LDA – linear discriminant analysis

LGD – loss given default

LR – linear regression

LSSVM – least squares support vector machine

LTV – loan to value

MARS – Multivariate adaptive regression splines

MDA – Multiple discriminant analysis

ML – machine learning

MLP – multilayer perceptron model

NN – neural network

NPL – non-performing loan

OLS – ordinary least squares

P2P – peer to peer

PD – probability of default

PLTR – penalised logistic tree regression

PTP – promise to pay

QDA – quadratic discriminant analysis

QMLE – quasi-maximum likelihood estimation

QR – quantile regression

QRNN – quantile regression neural network

RiR – Ridge regression

RF – random forest

RoR – Robust regression

RPC – right-party contact

RR – recovery rate

RT – regression tree

S&P – Standard & Poor's index

SMEs – small and medium-sized enterprises

SPV – special purpose vehicles

SVM – support vector machine

SVR – support vector regression

UR – Unemployment rate

XGBoost – Extreme gradient boosting

WOE – weight of evidence

Appendix B.- Summary of main LGD and RR model methodologies in literature

Table 7. Summary of main LGD and RR model methodologies in literature

Models methodologies	Authors
Linear regression (LR)	Li et al., 2023B; Dumitrescu et al., 2022; Ye & Bellotti, 2019; Nazemi & Fabozzi, 2018; Thomas & Bijak, 2015; Bijak, 2014; Hoechstetter & Nazemi, 2013; Thomas et al., 2012; Calabrese, 2012; Zhang & Thomas, 2012; Thomas et al., 2009; Chalupka & Kopecsni, 2008
Generalized linear model (GLM)	Beck et al., 2017; Han & Jang, 2013
Logistic regression	Xu et al., 2024; Dumitrescu et al., 2022; Calabrese & Zanin, 2022; Ye & Bellotti, 2019; Beck et al., 2017; Tanoue et al., 2017; Thomas & Bijak, 2015; Serrano-Cinca et al., 2015; Bijak, 2014; Hoechstetter & Nazemi, 2013; Loterman et al., 2012; Thomas et al., 2012; Calabrese, 2012; Hoechstetter et al., 2012; Matuszyk et al., 2010; Thomas et al., 2009
Ordinary least squares (OLS)	Jacobs, 2024; Nazemi et al., 2022; Li et al., 2022; Bellotti et al., 2021; Xia et al., 2021; Li et al., 2021; Tanoue et al., 2017; Han & Jang, 2013; Hoechstetter & Nazemi, 2013;

	Loterman et al., 2012; Bellotti & Crook, 2012; Khieu et al., 2012; Qi & Zhao, 2011
Logit-transformed linear regression, Logit-beta regression, and Local logit	Li et al., 2023A; Li et al., 2022; Sopitpongstorn et al., 2021; Thomas et al., 2009
Fractional response and logit regression	Kellner et al., 2022; Sopitpongstorn et al., 2021; Hoechstetter & Nazemi, 2013; Calabrese, 2012; Bastos, 2010; Qi & Zhao, 2011; Chalupka & Kopecsni, 2008
Beta regression (BR), Inflated beta regression, Joint beta regression, Beta mixture model, and Beta transformation/OLS (B-OLS)	Li et al., 2023A; Calabrese & Zanin, 2022; Kellner et al., 2022; Ye & Bellotti, 2019; Tanoue et al., 2017; Hoechstetter & Nazemi, 2013; Loterman et al., 2012; Thomas et al., 2012; Calabrese, 2012; Thomas et al., 2009
Box-Cox transformation/OLS (BC-OLS)	Hoechstetter & Nazemi, 2013; Thomas et al., 2009
Ridge regression (RiR)	Li et al., 2022; Bellotti et al., 2021; Xia et al., 2021; Nazemi & Fabozzi, 2018; Hoechstetter & Nazemi, 2013; Loterman et al., 2012
Robust regression (RoR)	Loterman et al., 2012
LASSO regression	Li et al., 2022; Bellotti et al., 2021; Li et al., 2021; Nazemi & Fabozzi, 2018
Tobit regression	Li et al., 2023A; Tanoue et al., 2017; Thomas & Bijak, 2015; Bijak, 2014; Hoechstetter & Nazemi, 2013; Bellotti & Crook, 2012
Principal components regression	Nazemi & Fabozzi, 2018
Quantile regression (QR)	Kellner et al., 2022
Multivariate adaptive regression splines (MARS)	Bellotti et al., 2021; Hoechstetter & Nazemi, 2013; Loterman et al., 2012
Support vector machine (SVM)	Xu et al., 2024; Li et al., 2023B; Dumitrescu et al., 2022; Bellotti et al., 2021; Hoechstetter & Nazemi, 2013; Hoechstetter et al., 2012; Chen et al., 2009
Least squares support vector machine (LSSVM)	Nazemi et al., 2022; Hoechstetter & Nazemi, 2013; Loterman et al., 2012
Support vector regression (SVR)	Nazemi et al., 2022; Xia et al., 2021; Li et al., 2021; Nazemi & Fabozzi, 2018
Survival analysis	Thomas & Bijak, 2015; Serrano-Cinca et al., 2015; Bijak, 2014; Hoechstetter & Nazemi, 2013; Zhang & Thomas, 2012
Linear discriminant analysis (LDA) and Quadratic discriminant analysis (QDA)	Xu et al., 2024
Multiple discriminant analysis (MDA)	Altman, 1968

WOE approach	Thomas et al., 2009
Generalised extreme value (GEV)	Calabrese & Zanin, 2022
Quasi-maximum likelihood estimation (QMLE)	Khieu et al., 2012
Decision tree (DT)	Li et al., 2024; Li et al., 2023A; Li et al., 2021; Thomas & Bijak, 2015; Bijak, 2014; Hoechstetter & Nazemi, 2013; Bellotti & Crook, 2012; Matuszyk et al., 2010
Classification and regression trees (CART)	Xu et al., 2024; Nazemi et al., 2022; Kellner et al., 2022; Bellotti et al., 2021; Sopitpongstorn et al., 2021; Xia et al., 2021; Nazemi & Fabozzi, 2018; Thomas & Bijak, 2015; Bijak, 2014; Hoechstetter & Nazemi, 2013; Loterman et al., 2012; Hoechstetter et al., 2012; Qi & Zhao, 2011; Bastos, 2010
Penalised logistic tree regression (PLTR)	Dumitrescu et al., 2022
Conditional inference trees	Bellotti et al., 2021
Chi-square automatic interaction detector (CHAID)	Hoechstetter et al., 2012
K-nearest neighbours (KNN)	Xu et al., 2024; Bellotti et al., 2021; Hoechstetter et al., 2012
Artificial neural networks (ANN)	Li et al., 2024; Li et al., 2023B; Nazemi et al., 2022; Dumitrescu et al., 2022; Kellner et al., 2022; Bellotti et al., 2021; Sopitpongstorn et al., 2021; Li et al., 2021; Hoechstetter & Nazemi, 2013; Loterman et al., 2012; Hoechstetter et al., 2012; Qi & Zhao, 2011
Deep neural network (DNN)	Nazemi et al., 2022
Quantile regression neural network (QRNN)	Kellner et al., 2022
Dmneural network	Hoechstetter et al., 2012
Gaussian processes	Calabrese & Zanin, 2022; Kellner et al., 2022; Bellotti et al., 2021; Sopitpongstorn et al., 2021; Hoechstetter & Nazemi, 2013; Qi & Zhao, 2011
Gaussian naive bayes (GNB) and Bernoulli naive bayes (BNB)	Xu et al., 2024
Random forest (RF)	Xu et al., 2024; Li et al., 2024; Li et al., 2023B; Dumitrescu et al., 2022; Li et al., 2022; Bellotti et al., 2021; Xia et al., 2021; Li et al., 2021
Boosting (gradient boosting, gradient boosting decision trees-GBDT, extreme gradient	Xu et al., 2024; Li et al., 2024; Nazemi et al., 2022; Li et al., 2022; Bellotti et al., 2021; Xia et

boosting-XGBoost, light gradient boosting-LightGBM, CatBoost, and Adaptive boosting-AdaBoost)	al., 2021; Li et al., 2021; Shoghi, 2019; Nazemi & Fabozzi, 2018
Bagging	Bellotti et al., 2021; Nazemi & Fabozzi, 2018
Cubist	Bellotti et al., 2021
Multilayer perceptron model (MLP)	Xu et al., 2024; Xia et al., 2021
Competing risks proportional hazards (CRPH)	Li et al., 2023B

Source: Own elaboration

Appendix C.- Summary of macroeconomic variables used in the literature and their effect on models

Table 8. Summary of macroeconomic variables used in the literature and their effect on models

Authors	Macroeconomic variables	Improves prediction flag
Wu et al., 2024	Consumer price index	Yes
Xu et al., 2024	Stock index, consumer price index, purchasing managers index, monetary aggregates growth	Yes
Galow et al., 2024	GDP, unemployment rate, stock returns, house prices, and long-term interest rates	Yes
Li et al., 2023A	Prime interest rates, producer price index, unemployment rate, and average real wage index	Yes
Li et al., 2023B	Lending interest rate, average wage growth, unemployment, producer price index, and industrial production index	Yes
Nazemi et al., 2022	Unemployment rate and excessive indebtedness rate	No
Kellner et al., 2022	20 variables for the USA loans and 18 for the Europe loans	Yes, but limited impact
Xia et al., 2021	3-month Treasury Bill rate, unemployment rate, average hourly earnings, and consumer confidence index	Yes
Bellotti et al., 2021	More than 400 macroeconomic, financial, structural, and housing market indicators from the European System of Central Banks (ESCB)	Yes
Li et al., 2021	S&P index	Yes
Wang et al., 2020	GDP, Industry median stock returns, and Credit spread	Yes

Radivojević et al., 2019	GDP, unemployment rate, inflation rate, household finale consumption expenditure, banks capital to assets ratio, and lending interest rate	Yes
Nazemi & Fabozzi, 2018	179 variables	Yes
Beck et al., 2017	GDP and unemployment rate	Yes, but limited impact
Han & Jang, 2013	GDP	Yes, but limited impact
Calabrese, 2012	GDP and unemployment rate	Yes
Bellotti & Crook, 2012	Bank interest rates, unemployment rate, and earnings growth (whole economy)	Yes
Khieu et al., 2012	GDP	Yes
Bruche & González-Aguado, 2010	GDP and credit cycle	Yes
Höcht & Zagst, 2010	GDP, TED spread, S&P index, Dow Jones Euro Stoxx index, growth rate in industrial production, volatility index from CBOE, VSTOXX and VDAX, treasury constant maturity rate, Euribor rate, consumer and price index, and S&P corporate default rate	Yes
Dermine & De Carvalho, 2006	GDP	No
Grunert & Weber, 2005	GDP, inflation rate, depreciations of banks quota, and interbank interest rate	Not confirmed
Friedman & Sandow, 2005	S&P-rated bonds default rate, S&P-rated bonds industry, and GDP	Yes
Schuermann, 2004	Business cycle moment: recession or expansion	Yes

Source: Own elaboration

Appendix D.- Summary of borrowers and debt variables in the literature

Table 9. Summary of borrowers and debt variables in the literature

Data source	Variables	Authors
Borrowers' characteristics	Natural and legal person	Kriebel & Yam, 2020; Beck et al., 2017; ECB, 2017; Hoechststoetter et al., 2012
	Age of the debtor/firm age	Xu et al., 2024; Li et al., 2024, Nazemi et al., 2022; Li et al., 2022; Bellotti et al., 2021; Kriebel & Yam,

		2020; Ye & Bellotti, 2019; Shoghi, 2019; Athreya et al., 2018; Thomas et al., 2012; Bellotti & Crook, 2012; Hoechstetter et al., 2012; Zhang & Thomas, 2012; Bastos, 2010; Thomas et al., 2009; Chalupka & Kopecsni, 2008; Dermine & De Carvalho, 2006; Grunert & Weber, 2005
	Income and expenditure of the debtor	Xu et al., 2024; Li et al., 2023A; Li et al., 2023B; Calabrese & Zanin, 2022; Xia et al., 2021; Li et al., 2021; Shoghi, 2019; Athreya et al., 2018; Serrano-Cinca et al., 2015; Bellotti & Crook, 2012; Zhang & Thomas, 2012
	Address, postal code, region/province, and country of residence/incorporation	Xu et al., 2024; Li et al., 2024; Nazemi et al., 2022; Li et al., 2022; Calabrese & Zanin, 2022; Bellotti et al., 2021; Kriebel & Yam, 2020; Shoghi, 2019; ECB, 2017; Calabrese, 2012; Bellotti & Crook, 2012; Hoechstetter et al., 2012; Höcht & Zagst, 2010; Chen et al., 2009; Grunert & Weber, 2005
	Years at current address	Bellotti & Crook, 2012; Zhang & Thomas, 2012; Matuszyk et al., 2010
	Residential status or asset presence	Xu et al., 2024; Li et al., 2023A; Li et al., 2023B; Nazemi et al., 2022; Li et al., 2022; Calabrese & Zanin, 2022; Serrano-Cinca et al., 2015; Thomas et al., 2012; Bellotti & Crook, 2012; Hoechstetter et al., 2012; Zhang & Thomas, 2012; Thomas et al., 2009
	Gender	Xu et al., 2024; Li et al., 2024; Li et al., 2022; Bellotti et al., 2021; Kriebel & Yam, 2020; Ye & Bellotti, 2019; Hoechstetter et al., 2012
	Marital status	Xu et al., 2024; Li et al., 2022; Bellotti et al., 2021; Ye & Bellotti, 2019; Zhang & Thomas, 2012
	Educational background	Xu et al., 2024; Li et al., 2022; Athreya et al., 2018
	Employment status, reference, experience, size or location	Xu et al., 2024; Li et al., 2023B; Calabrese & Zanin, 2022; Bellotti et al., 2021; Xia et al., 2021; Kriebel & Yam, 2020; Ye & Bellotti, 2019; Athreya et al., 2018; Serrano-Cinca

		et al., 2015; Bellotti & Crook, 2012; Zhang & Thomas, 2012
	Credit score, Credit bureau score, Agencies rating, Application score, PD, and internal scores, ratings and grades	Xu et al., 2024; Li et al., 2024; Li et al., 2023A; Li et al., 2023B; Nazemi et al., 2022; Bellotti et al., 2021; Xia et al., 2021; Li et al., 2021; Kriebel & Yam, 2020; Ye & Bellotti, 2019; Shoghi, 2019; Cheng & Cirillo, 2018; Tanoue et al., 2017; Serrano-Cinca et al., 2015; Han & Jang, 2013; Hoechstetter & Nazemi, 2013; Thomas et al., 2012; Bellotti & Crook, 2012; Hoechstetter et al., 2012; Khieu et al., 2012; Bastos, 2010; Matuszyk et al., 2010; Thomas et al., 2009; Chen et al., 2009; Dermine & De Carvalho, 2006; Bailey, 2002; Gupton, 2000
	Time or credit history with bank/creditor	Li et al., 2023A; Li et al., 2021; Shoghi, 2019; Bellotti & Crook, 2012; Zhang & Thomas, 2012; Bastos, 2010; Dermine & De Carvalho, 2006; Grunert & Weber, 2005; Bailey, 2002
	Credit history/inquiries	Xu et al., 2024; Li et al., 2023A; Li et al., 2023B; Calabrese & Zanin, 2022; Li et al., 2021; Serrano-Cinca et al., 2015; Hoechstetter et al., 2012; Khieu et al., 2012; Bailey, 2002
	Level of risk and Customer outlook	ECB, 2017; Bailey, 2002
	Borrowers' indebtedness, affordability, or viability	Li et al., 2023A; ECB, 2017; Serrano-Cinca et al., 2015; Zhang & Thomas, 2012; Höcht & Zagst, 2010; Grunert & Weber, 2005
	Presence and number of investors	Xu et al., 2024; Xia et al., 2021
	Borrower group, lending group, or consumer family, and number of members	Li et al., 2022; Xia et al., 2021; Calabrese, 2012; Zhang & Thomas, 2012; Matuszyk et al., 2010
	Firm size, age, and status	Wang et al., 2020; Han & Jang, 2013; Khieu et al., 2012; Höcht & Zagst, 2010; Chen et al., 2009; Grunert & Weber, 2005
	Firm industry and industry characteristics	Xu et al., 2024; Jacobs, 2024; Li et al., 2022; Kellner et al., 2022; Hoechstetter & Nazemi, 2013;

		Khieu et al., 2012; Qi & Zhao, 2011; Bastos, 2010; Bruche & González-Aguado, 2010; Höcht & Zagst, 2010; Chalupka & Kopecsni, 2008; Dermine & De Carvalho, 2006; Grunert & Weber, 2005; Schuermann, 2004; Gupton, 2000
	Firm legal form	Kellner et al., 2022; Kriebel & Yam, 2020; Höcht & Zagst, 2010; Grunert & Weber, 2005
	Firm financial ratios	Kriebel & Yam, 2020; Wang et al., 2020; Han & Jang, 2013; Khieu et al., 2012; Grunert & Weber, 2005; Altman, 1968
Debt or account characteristics	Arrears days past due and default date	Li et al., 2024; Nazemi et al., 2022; Bellotti et al., 2021; Xia et al., 2021; Kriebel & Yam, 2020; Wang et al., 2020; ECB, 2017; Bijak & Thomas, 2015; Han & Jang, 2013; Hoechstoeffer & Nazemi, 2013; Thomas et al., 2012; Bellotti & Crook, 2012; Zhang & Thomas, 2012; Matuszyk et al., 2010; Höcht & Zagst, 2010; Thomas et al., 2009; Chalupka & Kopecsni, 2008; Bailey, 2002
	Type of product	Nazemi et al., 2022; Kellner et al., 2022; Bellotti et al., 2021; Wang et al., 2020; Ye & Bellotti, 2019; Shoghi, 2019; Hoechstoeffer & Nazemi, 2013; Zhang & Thomas, 2012; Khieu et al., 2012; Qi & Zhao, 2011; Höcht & Zagst, 2010; Chen et al., 2009; Chalupka & Kopecsni, 2008; Bailey, 2002; Gupton, 2000
	Product/loan amount or EAD	Xu et al., 2024; Li et al., 2024; Jacobs, 2024; Li et al., 2023A; Li et al., 2023B; Nazemi et al., 2022; Li et al., 2022; Calabrese & Zanin, 2022; Bellotti et al., 2021; Xia et al., 2021; Li et al., 2021; Kriebel & Yam, 2020; Wang et al., 2020; Ye & Bellotti, 2019; Shoghi, 2019; Cheng & Cirillo, 2018; Beck et al., 2017; Tanoue et al., 2017; ECB, 2017; Serrano-Cinca et al., 2015; Bijak & Thomas, 2015; Han & Jang, 2013; Hoechstoeffer &

		Nazemi, 2013; Thomas et al., 2012; Calabrese, 2012; Bellotti & Crook, 2012; Hoechstetter et al., 2012; Khieu et al., 2012; Bastos, 2010; Matuszyk et al., 2010; Höcht & Zagst, 2010; Thomas et al., 2009; Chalupka & Kopecsni, 2008; Dermine & De Carvalho, 2006; Grunert & Weber, 2005; Bailey, 2002
	Collateral or personal guarantee, type and location of collateral, or product/collateral seniority	Jacobs, 2024; Li et al., 2022; Kellner et al., 2022; Wang et al., 2020; Tanoue et al., 2017; ECB, 2017; Hoechstetter & Nazemi, 2013, Calabrese, 2012; Khieu et al., 2012; Qi & Zhao, 2011; Bastos, 2010; Höcht & Zagst, 2010; Chen et al., 2009; Chalupka & Kopecsni, 2008; Dermine & De Carvalho, 2006; Grunert & Weber, 2005; Friedman & Sandow, 2005; Schuermann, 2004; Gupton, 2000
	Presence or number of multiple loans, products, and/or multiple creditors	Xu et al., 2024; Jacobs, 2024; Li et al., 2023A; Li et al., 2023B; Nazemi et al., 2022; Li et al., 2022; Calabrese & Zanin, 2022; Li et al., 2021; Kriebel & Yam, 2020; Ye & Bellotti, 2019; Serrano-Cinca et al., 2015; Bellotti & Crook, 2012; Zhang & Thomas, 2012; Höcht & Zagst, 2010; Chalupka & Kopecsni, 2008; Grunert & Weber, 2005; Bailey, 2002; Gupton, 2000
	Product age	Bellotti et al., 2021; Beck et al., 2017; Hoechstetter et al., 2012; Chalupka & Kopecsni, 2008
	Purpose of credit/loan	Xu et al., 2024; Li et al., 2023B; Calabrese & Zanin, 2022; Xia et al., 2021; ECB, 2017; Serrano-Cinca et al., 2015; Zhang & Thomas, 2012; Chen et al., 2009
	Product / loan interest rate	Xu et al., 2024; Li et al., 2024; Li et al., 2023A; Li et al., 2022; Calabrese & Zanin, 2022; Bellotti et al., 2021; Ye & Bellotti, 2019; Athreya et al., 2018; ECB, 2017; Bastos, 2010; Chalupka & Kopecsni, 2008; Dermine & De Carvalho, 2006

	Product / loan term	Xu et al., 2024; Calabrese & Zanin, 2022; Xia et al., 2021; Bijak & Thomas, 2015; Thomas et al., 2012; Zhang & Thomas, 2012; Thomas et al., 2009; Grunert & Weber, 2005
	Product / loan fees	Bellotti et al., 2021; Ye & Bellotti, 2019; Athreya et al., 2018
	Credit limit and limit utilization	Bellotti et al., 2021; Li et al., 2021; Ye & Bellotti, 2019; Höcht & Zagst, 2010; Bailey, 2002
	Revolving utilization and balance	Li et al., 2023A; Li et al., 2023B; Calabrese & Zanin, 2022; Xia et al., 2021; Li et al., 2021; Serrano-Cinca et al., 2015; Bailey, 2002
	Insurance presence	Athreya et al., 2018; Zhang & Thomas, 2012; Bailey, 2002
	LTV	Cheng & Cirillo, 2018; ECB, 2017; Hoechstetter & Nazemi, 2013; Chalupka & Kopecsni, 2008
	Re-structured, bankruptcy, hardship or insolvent cases	Li et al., 2023A; Li et al., 2023B; Kriebel & Yam, 2020; Wang et al., 2020; Beck et al., 2017; ECB, 2017; Khieu et al., 2012; Chen et al., 2009; Bailey, 2002
	Reason for default	Höcht & Zagst, 2010
	Currency	ECB, 2017
	Client portfolio	Nazemi et al., 2022

Source: Own elaboration

Appendix E.- Summary of repayment behaviour and C&R activities variables in the literature

Table 10. Summary of repayment behaviour and C&R activities variables in the literature

Variables	Authors
Days of centralized or outsourced management	Li et al., 2022; Kriebel & Yam, 2020; Shoghi, 2019; Höcht & Zagst, 2010
Presence or experience with external agencies	Li et al., 2022; Beck et al., 2017
The start of litigation process and its status	Li et al., 2022; Bailey, 2002
Trace or no trace of debtors	Matuszyk et al., 2010; Bailey, 2002
Duration and costs of the workout or resolution process	Jacobs, 2024; Tanoue et al., 2017; Chalupka & Kopecsni, 2008; Grunert & Weber, 2005; Bailey, 2002; Gupton, 2000

Credit bureau data	Xia et al., 2021; Ye & Bellotti, 2019; Bailey, 2002
Calls, contacts and visits (number, results and timings)	Bellotti et al., 2021; Ye & Bellotti, 2019; Shoghi, 2019; Bailey, 2002
Contact information (mobile phone, fixed-line phone, number of phones...)	Nazemi et al., 2022; Kriebel & Yam, 2020; Shoghi, 2019; Thomas et al., 2012; Thomas et al., 2009; Bailey, 2002
Type of bankruptcy process	Wang et al., 2020; Han & Jang, 2013
Repayment behaviour data (number of late payments, average number of late payments, differences of payments and cumulative differences, debt settlements)	Xu et al., 2024; Li et al., 2024; Li et al., 2023A; Nazemi et al., 2022; Calabrese & Zanin, 2022; Bellotti et al., 2021; Kriebel & Yam, 2020; Wang et al., 2020; Ye & Bellotti, 2019; Shoghi, 2019; Cheng & Cirillo, 2018; Beck et al., 2017; ECB, 2017; Thomas et al., 2016; Serrano-Cinca et al., 2015; Bijak & Thomas, 2015; Han & Jang, 2013; Bailey, 2002
Time between default and sell-off to third-party or date of sell-off	Nazemi et al., 2022; Hoechstetter et al., 2012

Source: Own elaboration

Appendix F.- General specifications for models construction

Table 11. General specifications for models construction

Tree levels used	7
Alpha for division	0.05
Alpha for melting	0.05
Epsilon for convergence	0.001
Maximum number of iterations for convergence	100
Using Bonferroni's correction	Yes
Chi-square method	Pearson
Minimum Parent Branch number of Records	2%
Minimum Child Branch number of Records	1%

Source: Own elaboration

Appendix G.- List of variables in the data sets

The final variables for each of the data sets were the eighty-one included in the following table.

Table 12. List of variables in the data sets

Variable	Description
1. "IUC"	Account identifier

2. "NUMERO"	Account identifier
3. "SERIE"	Account identifier
4. "Fecha"	Date: month and year
5. "IN01_Cartera_G"	Portfolio
6. "IN02_Vertical"	Vertical segment
7. "IN03_Originador"	Creditor
8. "IN04_Tipo_Producto"	Product type
9. "IN05_Saldo_INI_G"	Initial total debt amount
10. "IN06_Saldo_INI_INT_G"	Initial interest debt amount
11. "IN07_Saldo_INI_NOM_G"	Initial principal debt amount
12. "IN08_Saldo_INI_OTH_G"	Initial other debt amount
13. "IN09_Mes_Impago"	Month and year of the original debt
14. "IN10_Anio_Impago"	Year of the original debt
15. "IN11_Originador_G"	Creditor grouped
16. "IN12_Tipo_Producto_G"	Product type grouped
17. "IN13_Nacimiento_Min"	Oldest date of birth or date of company creation of participants
18. "IN14_Nacimiento_Max"	Youngest date of birth or date of company creation of participants
19. "IN15_NUM_TEL_FIJO_Sum"	Number of fixed line phone numbers
20. "IN16_NUM_TEL_MOVIL_Sum"	Number of mobile phone numbers
21. "IN17_NE_EXT"	Foreign nationality flag
22. "IN18_NE_NAC"	National nationality flag
23. "IN19_NE_OTH"	Other nationality flag
24. "IN20_Provincia"	Province
25. "IN21_TP_TIT"	Holder or active company flag
26. "IN22_TP_AVA"	Guarantor flag
27. "IN23_TP_OTH"	Other participant or relationship flag
28. "IN24_NumPersonas"	Number of participants
29. "IN25_TotalTelefonos"	Number of phone numbers
30. "IN26_CCAA"	Region
31. "IN27_Judicial"	Judicial flag
32. "IN28_Saldo_InicioCorte_G"	Monthly debt amount grouped
33. "IN29_SMS_L30D"	SMS sent in the last 30 days flag
34. "IN30_N_LLAMADAS_L30D"	Calls in the last 30 days flag
35. "IN31_RPC_L30D"	Right-party contact in the last 30 days flag
36. "IN32_PTP_L30D"	flag
37. "IN33_SMS_L90D"	Promise to pay in the last 30 days flag
38. "IN34_N_LLAMADAS_L90D"	SMS sent in the last 90 days flag
39. "IN35_RPC_L90D"	Calls in the last 90 days flag
40. "IN36_PTP_L90D"	Right-party contact in the last 90 days flag
41. "IN37_SMS_Ever"	flag
42. "IN38_N_LLAMADAS_Ever"	Promise to pay in the last 90 days flag
43. "IN39_RPC_Ever"	SMS sent ever flag
44. "IN40_PTP_Ever"	Calls ever flag
45. "IN41_Pago_L30D"	Right-party contact ever flag
46. "IN42_Pago_L90D"	Promise to pay ever flag

47. "IN43_Pago_Ever"	Payment in the last 30 days flag
48. "IN44_IPC_IndGeneral_VarAnual"	Payment in the last 90 days flag
49. "IN45_IPC_IndGeneral_VarAnual_TIL10"	Payment ever flag
50. "IN46_IPC_IndGeneral_VarMensual"	General consumer price index on annual variation
51. "IN47_IPC_IndGeneral_VarMensual_TIL10"	General consumer price index on annual variation grouped in deciles
52. "IN48_IPC_IndSubyacente_VarAnual"	General consumer price index on monthly variation
53. "IN49_IPC_IndSubyacente_VarAnual_TIL10"	General consumer price index on monthly variation grouped in deciles
54. "IN50_IPC_IndSubyacente_VarMensual"	Underlying consumer price index on annual variation
55. "IN51_IPC_IndSubyacente_VarMensual_TIL10"	Underlying consumer price index on annual variation grouped in deciles
56. "IN52_ICC"	Underlying consumer price index on monthly variation
57. "IN53_Ind_Sit_Actual"	Underlying consumer price index on monthly variation grouped in deciles
58. "IN54_Ind_Expectativas"	Consumer confidence index
59. "IN55_ICC_TIL10"	Consumer confidence index actual situation
60. "IN56_Ind_Sit_Actual_TIL10"	Consumer confidence index expectations
61. "IN57_Ind_Expectativas_TIL10"	Consumer confidence index grouped in deciles
62. "IN58_Paro_TasaActividad_VarAnual"	Consumer confidence index actual situation grouped in deciles
63. "IN59_Paro_TasaActividad_VarAnual_TIL10"	Consumer confidence index expectations grouped in deciles
64. "IN60_Paro_TasaParo_VarAnual"	Activity rate on annual variation
65. "IN61_Paro_TasaParo_VarAnual_TIL10"	Activity rate on annual variation grouped in deciles
66. "IN62_PIB_EmpleoEquivTiempoCompleto_VarAnual"	Unemployment rate on annual variation
67. "IN63_PIB_EmpleoEquivTiempoCompleto_VarAnual_TIL10"	Unemployment rate on annual variation grouped in deciles
68. "IN64_PIB_EmpleoEquivTiempoCompleto_VarTrim"	GDP full-time equivalent employment on annual variation
69. "IN65_PIB_EmpleoEquivTiempoCompleto_VarTrim_TIL10"	GDP full-time equivalent employment on annual variation grouped in deciles
70. "IN66_PIB_PreciosMercado_VarAnual"	GDP full-time equivalent employment on quarterly variation
	GDP full-time equivalent employment on quarterly variation grouped in deciles

71. "IN67_PIB_PreciosMercado_VarAnua I_TIL10"	GDP market prices on annual variation
72. "IN68_PIB_PreciosMercado_VarTrim"	GDP market prices on annual variation grouped in deciles
73. "IN69_PIB_PreciosMercado_VarTrim_T IL10"	GDP market prices on quarterly variation
74. "OUT01_Cob_60D"	GDP market prices on quarterly variation grouped in deciles
75. "OUT02_Cob_90D"	Amount paid in the following 60 days
76. "OUT03_Cob_180D"	Amount paid in the following 90 days
77. "OUT04_Pago_60D"	Amount paid in the following 180 days
78. "OUT05_Pago_90D"	Payment in the following 60 days flag
79. "OUT06_Pago_180D"	Payment in the following 90 days flag
80. "OUT07_Cobros_30D"	Payment in the following 180 days flag
81. "OUT08_Pago_30D"	Amount paid in the following 30 days Payment in the following 30 days flag

Source: Own elaboration

REFERENCES

Altman, E., Resti, A., & Sironi, A. (2004). Default recovery rates in credit risk modelling: a review of the literature and empirical evidence. *Economic notes*, 33(2), 183-208.

Altman, E. I., Brady, B., Resti, A., & Sironi, A. (2005). The link between default and recovery rates: Theory, empirical evidence, and implications. *The Journal of Business*, 78(6), 2203-2228.

Altman, E. I. (2021). Covid-19 and the credit cycle: 2020 revisited and 2021 outlook. *Journal of Credit Risk*.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, 23(4), 589-609.

Asociación Nacional de Empresas de Gestión de Créditos. (2024). *Estudio de mercado sobre el sector de la recuperación de deuda y la morosidad*. ANGECHO.

Athreya, K., Sánchez, J. M., Tam, X. S., & Young, E. R. (2018). Bankruptcy and delinquency in a model of unsecured debt. *International Economic Review*, 59(2), 593-623.

Axis Corporate. (2025). *Assets Under Management Report*. Edición VIII – 2025. Axis Corporate, Accenture.

Bailey, M. (2002). *Consumer Collections & Recoveries: Operations and Strategies*. White Box.

Balzac, H. (1827). *El arte de pagar las deudas y satisfacer a los acreedores sin desembolsar un céntimo, en diez lecciones*. Editorial Centellas.

Basle Committee on Banking Supervision, & Bank for International Settlements. (2005). *An explanatory note on the Basel II IRB risk weight functions*. Bank for International Settlements.

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

Bastos, J. A. (2014). Ensemble predictions of recovery rates. *Journal of Financial Services Research*, 46(2), 177-193.

Beck, T., Grunert, J., Neus, W., & Walter, A. (2017). What determines collection rates of debt collection agencies?. *Financial Review*, 52(2), 259-279.

- Bellotti, A., Brigo, D., Gambetti, P., & Vrms, F. (2021). Forecasting recovery rates on non-performing loans with machine learning. *International Journal of Forecasting*, 37(1), 428-444.
- Bellotti, T., & Crook, J. (2012). Loss given default models incorporating macroeconomic variables for credit cards. *International Journal of Forecasting*, 28(1), 171-182.
- Bijak, K. (2014). Performance measures of models to predict Loss Given Default: a critical review.
- Bijak, K., & Thomas, L. C. (2015). Modelling LGD for unsecured retail loans using Bayesian methods. *Journal of the Operational Research Society*, 66(2), 342-352.
- Bruche, M., & González-Aguado, C. (2010). Recovery rates, default probabilities, and the credit cycle. *Journal of Banking & Finance*, 34(4), 754-764.
- Calabrese, R. (2012). Predicting bank loan recovery rates with a mixed continuous-discrete model. *Applied stochastic models in business and industry*, 30(2), 99-114.
- Calabrese, R., & Zanin, L. (2022). Modelling spatial dependence for Loss Given Default in peer-to-peer lending. *Expert Systems with Applications*, 192, 116295.
- Carleo, A., & Rocci, R. (2024). Functional clustering of NPLs recovery curves. *Socio-Economic Planning Sciences*, 95, 102018.
- Chalupka, R., & Kopecsni, J. (2008). *Modelling bank loan LGD of corporate and SME segments: a case study* (No. 27/2008). IES Working Paper.

- Chen, H., Ma, Y.C., Chen, M. Z., Tang, Y., Wang, B., Chen, M., & Yang, X. G. (2009). Recovery Discrimination based on Optimized-Variables Support Vector Machine for Nonperforming Loan. *Systems Engineering-Theory & Practice*, 29(12), 23-30.
- Cheng, D., & Cirillo, P. (2018). A reinforced urn process modeling of recovery rates and recovery times. *Journal of Banking & Finance*, 96, 1-17.
- Coleman, A. M. (2004). *Collection management handbook: The art of getting paid*. John Wiley & Sons.
- Crespo, I., & Govindarajan, A. (2018). The analytics-enabled collections model. *McKinsey on payments. Special edition on Advanced Analytics in Banking*, 11(30), 4-12.
- Dermine, J., & De Carvalho, C. N. (2006). Bank loan losses-given-default: A case study. *Journal of Banking & Finance*, 30(4), 1219-1243.
- Drozdz, L. A., & Serrano-Padial, R. (2013). *Modeling the credit card revolution: the role of debt collection and informal bankruptcy* (No. 13-12). Federal Reserve Bank of Philadelphia.
- Dumitrescu, E., Hué, S., Hurlin, C., & Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research*, 297(3), 1178-1192.
- European Banking Authority. (2019). *EBA report on NPLs. Progress made and challenges ahead*. European Banking Authority.

European Central Bank. (2017). Guidance to banks on non-performing loans. *Guidance to Banks on Non-Performing Loans, Banking Supervision Division*, 1-131.

Fedaseyeu, V. (2015). *Debt collection agencies and the supply of consumer credit*. Research Department, Federal Reserve Bank of Philadelphia

Fedaseyeu, V. (2020). Debt collection agencies and the supply of consumer credit. *Journal of Financial Economics*, 138(1), 193-221.

Fedaseyeu, V., & Hunt, R. M. (2015). *The economics of debt collection: Enforcement of consumer credit contracts*. Research Department, Federal Reserve Bank of Philadelphia.

Fonseca, J., Strair, K., & Zafar, B. (2017). *Access to credit and financial health: evaluating the impact of debt collection* (No. 814, pp. 1-37). Federal Reserve Bank of New York.

Friedman, C. A., & Sandow, S. (2005). Estimating conditional probability distributions of recovery rates: a utility-based approach. Standard & Poor's Risk Solutions Group.

Galow, B., Georgescu, O. M., & Marques, A. P. (2024). *Loss-given-default and macroeconomic conditions* (No. 2954). ECB Working Paper.

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

Giovine, C., Lerner, L., Thomas, R., Kakulavarapu, S., Chung, V., & Motwani, Y. (2024). Extracting value from AI in banking: Rewiring the enterprise. *McKinsey & Company*.

Grunert, J., & Weber, M. (2005). Recovery rates of bank loans: empirical evidence for Germany.

Gupton, G. M., Gates, D., & Carty, L. V. (2000). Bank loan loss given default. *Moody's Investors Service, Global Credit Research, November*.

Gürtler, M., & Hibbeln, M. (2013). Improvements in loss given default forecasts for bank loans. *Journal of Banking & Finance*, 37(7), 2354-2366.

Han, C., & Jang, Y. (2013). Effects of debt collection practices on loss given default. *Journal of Banking & Finance*, 37(1), 21-31.

Higginson, M., Jacques, F., & Rudisuli, R. (2018). The seven pillars of (collections) wisdom. *McKinsey on Risk*, 5.

Höcht, S., & Zagst, R. (2010). Loan Recovery Determinants--A Pan-European Study. *Available at SSRN 2443724*.

Hoechstetter, M., Nazemi, A., Rachev, S. T., & Bozic, C. (2012). Recovery rate modelling of non-performing consumer credit using data mining algorithms. *Review of*.

Hoechstetter, M., & Nazemi, A. (2013). Analysis of loss given default. *Investment management and financial innovations*, (10, Iss. 4), 70-79.

Hunt, R. M. (2007). Collecting consumer debt in America. Available at SSRN 993249.

Informa, D.B.K. (2023). Empresas de gestión de impagados. Informes Especiales.

Jacobs, M. (2024). Modeling Ultimate Loss-Given-Default and Time-to-Resolution on Corporate Debt. Available at SSRN 4738268.

Jimenez, G., & Saurina, J. (2006). Credit cycles, credit risk, and prudential regulation. *International Journal of Central Banking*, 2(2).

Jonasson, T., Papaioannou, M. G., & Williams, M. (2019). Debt management. *Sovereign Debt: A Guide for Economists and Practitioners*, 192.

Kellner, R., Nagl, M., & Rösch, D. (2022). Opening the black box—Quantile neural networks for loss given default prediction. *Journal of Banking & Finance*, 134, 106334.

Khieu, H. D., Mullineaux, D. J., & Yi, H. C. (2012). The determinants of bank loan recovery rates. *Journal of Banking & Finance*, 36(4), 923-933.

Kriebel, J., & Yam, K. (2020). Forecasting recoveries in debt collection: Debt collectors and information production. *European Financial Management*, 26(3), 537-559.

Krumbein, W. C. (1924). Collection Agencies. *University Journal of Business*, 48-67.

Lawrence, D. B. (1992). *Handbook of consumer lending*. Englewood Cliffs, NJ, Prentice-Hall.

Li, A., Li, Z., & Bellotti, A. (2023A). Predicting loss given default of unsecured consumer loans with time-varying survival scores. *Pacific-Basin Finance Journal*, 78, 101949.

Li, G., Chen, M., & Yang, X. (2022, July). Loss Given Default Modeling and Variable Importance Analysis of Credit Loans. In *2022 41st Chinese control conference (CCC)* (pp. 7491-7496). IEEE.

Li, J., Kang, J., Wu, J., Wang, H., & Yang, X. (2024). Research on credit card default repayment prediction model. *The Journal of Finance and Data Science*, 10, 100136.

Li, K., Zhou, F., Li, Z., Yao, X., & Zhang, Y. (2021). Predicting loss given default using post-default information. *Knowledge-Based Systems*, 224, 107068.

Li, Z., Li, A., Bellotti, A., & Yao, X. (2023B). The profitability of online loans: A competing risks analysis on default and prepayment. *European Journal of Operational Research*, 306(2), 968-985.

Lin, Z. Y., & Lu, J. W. (2024). The impact of COVID-19 on global investor attention. *Pacific-Basin Finance Journal*, 88, 102522.

Loterman, G., Brown, I., Martens, D., Mues, C., & Baesens, B. (2012). Benchmarking regression algorithms for loss given default modeling. *International Journal of Forecasting*, 28(1), 161-170.

Malikidou, D., & Strohbach, W. (2025). Predicting Bank Distress in Europe: Using Machine Learning and a Novel Definition of Distress. *European Banking Authority Research Paper*, (21).

- Malinconico, A., & Virguti, E. (2024). Business Models and Best Practices in Debt Recovery Services: an overview of the Italian industry. In *Conference Proceedings*.
- Matuszyk, A., Mues, C., & Thomas, L. C. (2010). Modelling LGD for unsecured personal loans: Decision tree approach. *Journal of the Operational Research Society*, 61(3), 393-398.
- Molina, S. (2015). *El ciclo del crédito*. Editorial Almuzara.
- Nazemi, A., & Fabozzi, F. J. (2018). Macroeconomic variable selection for creditor recovery rates. *Journal of Banking & Finance*, 89, 14-25.
- Nazemi, A., & Fabozzi, F. J. (2024). Interpretable machine learning for creditor recovery rates. *Journal of Banking & Finance*, 164, 107187.
- Nazemi, A., Rezazadeh, H., Fabozzi, F. J., & Höchstötter, M. (2022). Deep learning for modeling the collection rate for third-party buyers. *International Journal of Forecasting*, 38(1), 240-252.
- O'Brien, L., Chen, V., Ramsay, I., & Ali, P. (2021). An Impending 'Avalanche': Debt Collection and Consumer Harm after COVID-19. *Australian Business Law Review*, 49(2), 84-115.
- Qi, M., & Zhao, X. (2011). Comparison of modeling methods for loss given default. *Journal of Banking & Finance*, 35(11), 2842-2855.
- Radivojević, N., Cvijanović, D., Sekulic, D., Pavlovic, D., Jovic, S., & Maksimović, G. (2019). Econometric model of non-performing loans determinants. *Physica A: Statistical Mechanics and its Applications*, 520, 481-488.

- Ramos González, M. R., Ureña, A. P., & Fernández-Aguado, P. G. (2023). Forecasting for regulatory credit loss derived from the COVID-19 pandemic: A machine learning approach. *Research in International Business and Finance*, 64, 101907.
- Repullo, R., Saurina, J., & Trucharte, C. (2010). Mitigating the Pro-cyclicality of Basel II. *Economic Policy*, 25(64), 659-702.
- Rosenberg, E., & Gleit, A. (1994). Quantitative methods in credit management: a survey. *Operations research*, 42(4), 589-613.
- Schuermann, T. (2004). What do we know about loss given default?
- Serrano-Cinca, C., Gutiérrez-Nieto, B., & López-Palacios, L. (2015). Determinants of default in P2P lending. *PloS one*, 10(10), e0139427.
- Shoghi, A. (2019). Debt collection industry: machine learning approach. *Journal of Money and Economy*, 14(4), 453-473.
- Sopitpongstorn, N., Silvapulle, P., Gao, J., & Fenech, J. P. (2021). Local logit regression for loan recovery rate. *Journal of Banking & Finance*, 126, 106093.
- Spuchlaková, E., & Cug, J. (2015). Credit risk and LGD modelling. *Procedia Economics and Finance*, 23, 439-444.
- Stifler, L., & Parrish, L. (2014). *Debt Collection & Debt buying. The State of Lending in America & its Impact on the U.S. Households*. Center for Responsible Lending.

Tanoue, Y., Kawada, A., & Yamashita, S. (2017). Forecasting loss given default of bank loans with multi-stage model. *International Journal of Forecasting*, 33(2), 513-522.

Thomas, L., & Bijak, K. (2015). *Impact of segmentation on the performance measures of LGD models*.

Thomas, L. C., Matuszyk, A., So, M. C., Mues, C., & Moore, A. (2016). Modelling repayment patterns in the collections process for unsecured consumer debt: A case study. *European Journal of Operational Research*, 249(2), 476-486.

Thomas, L. C., Matuszyk, A., & Moore, A. (2012). Comparing debt characteristics and LGD models for different collections policies. *International Journal of Forecasting*, 28(1), 196-203.

Thomas, L. C., Matuszyk, A., & Moore, A. (2009). Collections policy comparison in LGD modelling.

Várgedő, B. (2024). A stochastic loss given default model of Hungarian residential mortgages. *Acta Oeconomica*, 74(3), 299-327.

Wang, H., Forbes, C. S., Fenech, J. P., & Vaz, J. (2020). The determinants of bank loan recovery rates in good times and bad—new evidence. *Journal of Economic Behavior & Organization*, 177, 875-897.

Wu, Y., Arora, G., & Mei, X. (2024). Using CPI in Loss Given Default Forecasting Models for Commercial Real Estate Portfolio. *arXiv preprint arXiv:2402.15498*.

Xia, Y., Zhao, J., He, L., Li, Y., & Yang, X. (2021). Forecasting loss given default for peer-to-peer loans via heterogeneous stacking ensemble approach. *International Journal of Forecasting*, 37(4), 1590-1613.

Xu, Q., Liu, C., Luo, J., & Liu, F. (2024). Using machine learning to investigate the determinants of loan default in P2P lending: Are there differences between before and during COVID-19?. *Pacific-Basin Finance Journal*, 88, 102550.

Ye, H., & Bellotti, A. (2019). Modelling recovery rates for non-performing loans. *Risks*, 7(1), 19.

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