1 Introduction

1.1 Motivation and Objectives

Financial institutions play a crucial role in maintaining the stability of the financial system by ensuring the supply of money and credit and supporting the transfer of risk between entities. In general, a resilient banking system supports the real economy and contributes positively to sustainable economic growth (Basel Committee on Banking Supervision, 2017). However, in their role as intermediaries, financial institutions are exposed to different types of risk. Compared to other risks, such as market risk and operational risk, credit risk accounts for the largest share of financial institutions' regulatory capital requirements. For example, the European Banking Authority's latest Risk Assessment Report (2022) identifies that 83.3% of the risk-weighted assets of 131 major European (EU) banks were attributable to credit risk as of June 2021. Effective credit risk management is therefore of great importance not only for financial institutions but also for the economy in general, as it enables financial institutions to fulfill their important role as intermediaries at all times.

Credit risk is most simply defined as the possibility that a debtor will not meet its obligations according to the agreed terms (Basel Committee on Banking Supervision, 2000). It can be characterized by three risk parameters, which are generally modeled as stochastic variables. Probability of default (PD) refers to the probability that a borrower will default on its payment obligations in a given future period. Exposure at default (EAD) defines the outstanding amount at the time of default. Loss given default (LGD) specifies the proportion of the outstanding amount that will be lost due to the non-fulfillment of the obligation. The objective of credit risk parameter modeling is to provide financial institutions with a framework for assessing and managing the risk associated with their loan portfolios. Given their systemic importance, banks are subject to regulatory requirements for credit risk management and parameter modeling. Under the Basel frameworks (Basel II and III, see Basel Committee on Banking Supervision (2006, 2010, 2017)), banks are permitted to use the Internal Ratings Based (IRB) approach to quantify their capital requirements for credit risk. While proprietary statistical models are permitted to estimate PD under the foundation IRB approach, proprietary estimates of LGD and EAD are reserved for the advanced IRB approach. One of the Committee's objectives in adopting the IRB approach is to align capital requirements more closely with the level of credit risk to which a financial institution is exposed. Ideally, the use of proprietary statistical models should reduce banks' regulatory capital requirements and free up additional capital that could be used for other banking activities. However, accurate predictions of the risk parameters are important for other reasons as well. For instance, banks use the risk parameters to determine an

appropriate capital buffer that should absorb potential losses from their business failures (Basel Committee on Banking Supervision, 2017). This is important because holding a capital buffer allows banks to mitigate the negative impact of unpredictable risk events (such as economic crises) and to protect themselves from severe financial distress. It also enables banks to provide liquidity and lending even in difficult economic times. This is important because a decline in lending can prolong or exacerbate an economic downturn. In extreme cases, the reduction of intermediation by financial institutions can even cause a recession (Ivashina and Scharfstein, 2010). In addition, accurate predictions of the credit risk parameters are important for banks' internal risk management. They allow banks to differentiate between low-risk and high-risk borrowers and to adjust lending policies to minimize the risk of loss from borrower default. They are also crucial for pricing purposes and credit limit management. Inaccurate predictions can lead to loans being made on inappropriate terms, resulting in higher losses or lower profitability. Therefore, accurate predictions offer banks a competitive advantage, whereas weak predictions can lead to adverse selection.

In recent years, computational power and digital storage capacities have increased substantially, while costs have dropped sharply. This allows researchers and practitioners to use more sophisticated and computationally intensive machine learning algorithms in credit risk management (Federation of European Risk Management Associations, 2019). Machine learning is fundamentally changing the modeling paradigm, moving from basic statistical methods to advanced learning algorithms that can make accurate predictions based on even highly nonlinear and complex data. Surveys of the Bank of England (2019) and Deutsche Bundesbank (2020) indicate that machine learning applications are gradually being used in banking practice. While real-world applications already exist in some institutions, most potential use cases are expected in the coming years (Bank of England, 2019). Therefore, many financial institutions are still in the early stages of adoption. From a statistical perspective, machine learning algorithms are particularly well suited to deal with typical characteristics found in banking data. For example, credit portfolios often exhibit complex data structures and typically involve large datasets in terms of the number of loans and explanatory variables. In addition, these datasets typically contain quantitative and qualitative variables. Missing data and outliers are also common. Machine learning algorithms can easily cope with all these difficulties, giving them great potential for credit risk modeling.

This thesis sheds light on the application of statistical and machine learning methods to LGD modeling. Most LGD studies are conducted in the spirit of "horse races", where different statistical and machine learning methods are compared in their predictive performances to determine the superior method (see, e.g., Qi and Zhao (2011), Loterman et al. (2012), or Kaposty et al. (2020)). In summary, views on how well various LGD estimation methods perform are

mixed, raising doubts as to which method is best. For example, while Yao et al. (2017) examine data on bank credit cards from United Kingdom (UK) and conclude that a combination of least squares support vector regression and ordinary least squares regression leads to the best out-of-sample estimation accuracy, Hurlin et al. (2018) base their analysis on defaulted customers in Brazil and find that the random forest mostly outperforms other methods. In particular, the different results can be attributed to the different countries where the considered loan portfolios are located (see, e.g., Bastos (2010)). More specifically, most studies focus on a country-specific loan portfolio, but there may be national differences in bankruptcy law or borrower characteristics that affect the performance of the LGD models (see, e.g., Grunert and Weber (2009)). Moreover, studies indicate that the LGD distributions in credit portfolios seem to differ between regions and countries (see, e.g., Grippa et al. (2005)). Thus, it remains unclear which characteristics of an LGD distribution are responsible for the different performance results in the literature. Against this background, the first objective of this thesis is to identify the distributional features relevant to the quality of LGD estimation methods, and subsequently, determine the methods that have the highest estimation accuracy for the relevant distribution types.

To date, the question of the best LGD method remains unanswered, but there is a general consensus that machine learning algorithms outperform traditional methods such as linear regression. However, the main limitations of machine learning algorithms come from their lack of explainability and interpretability. They are often referred to as "Black Boxes" because their estimation process cannot be easily explained, which severely limits their use in credit risk modeling, especially with regard to regulatory requirements (see, e.g., European Banking Authority (2013a)). Some techniques from the field of Explainable Artificial Intelligence (XAI) can be used to make predictions of machine learning models more understandable (see, e.g., Bussmann et al. (2021) and Bastos and Matos (2022)), but they have numerous limitations and carry an increased risk of misinterpretation. For example, Kaposty et al. (2020) promote the use of variable permutation, which measures the importance of a variable by calculating the change in the model's estimation error after permuting the variable. However, this method leads to biased error measures if the variable to be permuted is correlated with other variables. At the time of writing, there is no clear consensus on whether XAI techniques can adequately explain machine learning models, as reflected in regulatory publications (see, e.g., European Commission (2020) or European Banking Authority (2020)). As a result, machine learning is rarely used in the credit risk modeling industry, and linear regression remains the main model for LGD modeling due to its simplicity and intrinsic interpretability. However, the main challenge in using regression models is to exogenously determine the best subset of variables to include in the regression model. The wrong choice of variables can lead to problems such as biased regression coefficients and a reduction in out-of-sample predictive accuracy. Calibrating a regression model is therefore particularly difficult when a large number of variables are available and more complex terms,

such as variable interactions, are to be considered. Against this background, the second objective of this thesis is to optimize the calibration process of linear regression in such a way that its prediction accuracy is increased and comparable to that of machine learning models. Accordingly, the result should be an optimized regression model that meets the regulatory requirements for high accuracy and transparency.

Apart from the comparative LGD studies, it has been pointed out that a single statistical model may not be sufficient to capture the risk characteristics of different individuals in a credit portfolio (see, e.g., Bakoben et al. (2020)). Clustered modeling can be used to overcome this problem. In clustered modeling, borrowers are segmented based on their similarities through cluster analysis, and a separate LGD model is developed for each cluster, resulting in increased predictive accuracy. Unambiguously, its effectiveness depends on the quality of the segmentation, which in turn depends primarily on the choice of variables used in the cluster analysis. However, the optimal choice of variables for clustering is a major challenge, especially for high-dimensional credit data. An incorrect choice can lead to overlapping, indistinguishable, and uninformative clusters (Fop and Murphy, 2018), which negatively affects the predictive performance in separate modeling. In addition, high-dimensional credit data can be meaningfully clustered in many ways; that is, it is not necessary to identify the variables that lead to the best clustering, but those that enable the best prediction of the LGDs in separate modeling. The literature has already addressed the challenge of variable selection (see, e.g., Caruso et al. (2021) or Yuan et al. (2022)), but all proposed basic approaches have substantially limitations and/or are inappropriate for regulatory disclosure reasons. For instance, most studies use principal component analysis (PCA) for variable selection (see, e.g., Yoshino and Taghizadeh-Hesary (2019) and Le et al. (2021)). PCA selects variables by reducing the dimensionality of the data; that is, it creates new informative variables as linear combinations or mixtures of the original variables. In this way, variables are automatically selected for clustering, but at the cost of a lower understanding of meaning. However, as mentioned earlier, regulators generally require explainability in credit risk modeling, which limits the practicality of PCA as a variable selection technique. Against this background, the third objective of this thesis is to develop an intelligent variable selection procedure that meets regulatory requirements and can be used to determine an optimal set of variables for clustering in clustered modeling.

In summary, there are several shortcomings in LGD modeling in the financial literature. First, it remains unclear which features of an LGD distribution are responsible for the different performance results of LGD methods in the literature. As a result, there is no consensus on which method is most appropriate for modeling. Second, there is a trade-off between transparency and accuracy in LGD models. More complex machine learning models may provide better predictive performances at the cost of less explainability and comprehension of the model's

functioning. Conversely, linear regression offers high interpretability but seemingly limited predictive accuracy. Since regulators prefer to limit the intrinsic complexity of LGD models, the best way to resolve this trade-off is to improve the predictive accuracy of interpretable linear regression. Third, variable selection in clustered modeling is a challenge that needs to be addressed. Against this background, the following important research questions can be formulated:

- Which distributional characteristics are relevant to the quality of LGD estimation methods? Which methods have the highest estimation accuracy for relevant distribution types?
- How can the predictive performance of interpretable linear regression be improved? Can linear regression be optimized to compete with machine learning methods?
- How to determine an optimal set of variables for clustering in clustered modeling to achieve a particularly high prediction quality in the individual segments?

To answer these questions, a unique international dataset from Global Credit Data (GCD) of resolved defaulted loans from small- and medium-sized enterprises (SMEs) and large corporations (LCs) is used.[1] For the analyses in chapters 3-5, subsets of this dataset are used in the following ways: First, using 32,851 defaulted loans by SMEs from 16 European countries during 2000-2016, chapter 3 identifies heterogeneities among LGD distributions through cluster analysis. The analysis leads to three clusters, whose distributions essentially differ in their modality type. More specifically, a (nearly) symmetric bimodal distribution, an asymmetric (positively skewed) bimodal distribution, and a (positively skewed) unimodal distribution are found. For each modality type, the estimation accuracies of 20 different methods[2] are tested based on their out-of-sample performances. It is shown that the specific modality type is crucial for the best method.

Second, in chapter 4, linear regression is augmented with an automated intelligent variable selection process that is optimized using machine learning techniques. The effectiveness of the optimized regression model is investigated in a Monte Carlo experiment and empirical analysis using 9,457 defaulted loans of small, medium, and large enterprises from the United States (US) over the period 2000-2019. It is shown that linear regression with the optimal variable set can predict credit risk significantly more accurately than regressions using standard variable selection techniques and is competitively comparable with the best machine learning methods.

Third, chapter 5 proposes a clustered modeling approach in which variable selection for clustering is optimized using machine learning models. Under this approach, variables that contain relevant information for predicting credit risk are automatically and effectively identified and used in the

[1] Chapter 2 provides a brief overview of the data provider and the data itself.
[2] See Chapter 2 for a description of the methods.

cluster analysis, thereby considerably reducing the risk of creating uninformative clusters. The superiority of the optimized clustered approach is investigated through an empirical analysis using the same dataset as in chapter 4. It is demonstrated that the optimized clustered approach outperforms non-clustered modeling and clustered approaches using basic variable selection methods.

1.2 Course of Investigation

In order to analyze the research questions mentioned above, the course of investigation is as follows.

Chapter 2 introduces the GCD database and provides a basic understanding of LGD estimation. The chapter begins with a brief overview of the institutional background of the data provider and the data itself. The datasets used in the empirical studies are subsamples of the GCD database and are described separately in each chapter. Section 2.2 describes the theoretical background of the LGD estimation methods used in the comparative analyses and presents the procedure for determining appropriate hyperparameter values for the competing methods. Section 2.3 provides the measures used to compare the predictive performances of the methods.

Chapter 3 identifies heterogeneities among LGD distributions through cluster analysis and determines the LGD methods that have the highest estimation accuracy for the relevant distribution types. The chapter begins with a recap of the fundamentals and the research questions to be answered in this chapter. Section 3.2 introduces the empirical data, provides descriptive statistics, and explains the LGD estimation methods used in the empirical analysis. Cluster analysis is then performed to identify heterogeneities among credit risk parameter distributions. Section 3.3 presents the procedure used to compare the predictive performances of the LGD methods. Subsequently, the selected hyperparameter values for the competing methods in each cluster are presented. Finally, the empirical analysis is conducted and the cluster-specific results of the comparative analysis are presented and discussed. In Section 3.4 the robustness of the results is tested. The interim results of this chapter are summarized in Section 3.5.

In Chapter 4 linear regression is augmented with an automated intelligent variable selection process that is optimized using machine learning techniques. The chapter begins with an overview of the relevant literature and the motivation for the analysis. In Section 4.2, the variable selection model based on arbitrary machine learning algorithms is introduced. Section 4.3 analyzes the ability of optimized linear regression to correctly capture the linear and nonlinear effects occurring in simulated data through a Monte Carlo experiment. Section 4.4 introduces the empirical data and shows its plausibility using descriptive statistics. Next, the settings

for the comparative analyses and the competing models are described. Finally, the procedure and measures for comparing out-of-sample model performances are explained. The empirical analysis is conducted in Section 4.5. Here, the results of the procedure for selecting the variables to obtain an optimized regression model are presented. In addition, the in-sample estimation results of the optimized model are shown and variable effects are discussed. Finally, the out-of-sample performances of all models are compared using several evaluation criteria. In Section 4.6, several robustness checks are performed. Section 4.7 concludes the chapter.

Chapter 5 improves the predictive accuracy of the clustered LGD modeling approach by addressing the challenge of variable selection in clustering. Specifically, it proposes an intelligent variable selection process that is optimized using machine learning. The chapter begins with a review of the fundamentals and research questions. In Section 5.2, the optimized clustered approach based on arbitrary machine learning algorithms for variable selection is introduced. Section 5.3 presents the empirical framework, that is, the empirical data and settings used for the comparative analyses and the competing modeling approaches. Section 5.4 performs the empirical analysis. Here, the results of determining the variable importance in gradient-boosted trees using Shapley values are presented. Next, the clustering results of all the competitive clustering models are shown, and those of the optimized clustering model are described in more detail. Finally, the results of the comparative out-of-sample analyses are presented and evaluated. In Section 5.5 the robustness of the results is tested. Section 5.6 summarizes the results obtained in this chapter.

Chapter 6 concludes this thesis.

2 Data and LGD Estimation

2.1 Data

Established in 2004, GCD[3] is a global association of banks specializing in credit risk data and analysis. It was originally founded as a credit data pooling initiative to help member banks prepare for Basel II and achieve the advanced IRB status. Today, the GCD provides the world's largest database for credit risk modeling and is internationally recognized as the standard for credit data collection. Membership has grown from 11 to currently over 50 banks, and the geographic coverage of the GCD databases, originally limited to Europe, has been expanded to include banks in Africa, Australia and North America (see Figure 2.1). The primary objective of GCD is to enhance the risk management capabilities of its member banks. Through secure and confidential data sharing, the GCD facilitates the collection and anonymization of credit risk data provided by its members. This data serves as the basis for benchmarking and enables participants to gain insight into credit risk trends, default rates, and portfolio performance metrics. Over the years, the GCD has continuously professionalized and enhanced its databases to ensure that the use of its data meets regulatory requirements. The database is updated semi-annually with new defaults from member banks.

The analyses in the following chapters are based on GCD's dynamic and growing default database. At the time of writing, the raw dataset contains more than 300,000 individual loans with a total exposure of more than €750,000 billion from more than 70,000 obligors in 120 different countries. The sample period spans from January 1990 to December 2022. The subsamples used in the empirical analyses are adapted to the research questions using different filtering rules, which are explained separately in each chapter. Panel A of Table 2.1 reports the number of loans and defaults with corresponding exposures for different facility asset classes. In total, there are more than 150,000 defaults on individual loans, with small and medium-sized enterprises and large corporates accounting for the largest share of defaults at approximately 81%. Because these two asset classes are categorized as general corporate exposures under regulatory guidelines (see, for example, Basel Committee on Banking Supervision (2017)), they are used in the empirical analyses. The other specialized lending asset classes are not considered. Panel B of Table 2.1 shows the share of defaulted loans by region. Europe and North America account for the largest shares. Therefore, the analyses focus on these regions. For each loan in the database, various information is available at the time of default. This includes, for example, the EAD, the number of collateral and guarantees, and the seniority. In addition, macroeconomic

[3] For further information, see `https://www.globalcreditdata.org/`.

Figure 2.1: Location of the GCD member banks

data such as gross domestic product (GDP) and unemployment rates are included in the empirical analyses. For a detailed description of the explanatory variables, see Table A.1 in the Appendix.

For the sake of clarity and comprehensibility, some important terms need to be clarified. Borrowers in the small and medium corporate asset class are defined in §218 and §273 of the Basel II Accord as having reported sales for the consolidated group to which the firms belongs of less than €50 million. For large borrowers, the reported sales for the consolidated group to which the firm belongs are greater than or equal to €50 million. In addition, a default is considered to have occurred with respect to a particular obligor when one or both of the following two events have occurred (Basel Committee on Banking Supervision, 2006). First, the bank believes that the obligor is unlikely to pay its credit obligations to the banking group in full, without recourse by the bank to actions such as the realization of collateral. Second, the obligor is more than 90 days past due on any material credit obligation to the banking group. Overdrafts are considered past due when the customer has exceeded an advised limit or has been advised of a limit that is less than the current outstandings. LGDs are calculated using workout recovery rates, which are the difference between all discounted post-default incoming cash flows (F^+) and all discounted post-default costs (C^-), divided by the exposure at default. That is,

$$LGD = 1 - \frac{\sum F^+ - \sum C^-}{EAD}. \tag{2.1}$$

Incoming cash flows comprise principal and interest payments, recorded book value of collateral, received fees, and commissions. Costs include legal expenses, administrator and receiver fees, liquidation expenses, and other external workout costs. All cash flows are discounted at either the three-month London Interbank Offered Rate (LIBOR) or the three-month Euro Interbank Offered Rate (EURIBOR) of the respective default date, depending on whether US or EU data is used in the empirical analysis.

Table 2.1: Summary statistics (GCD database)

Panel A: LGD 2023 Datapool overview by facility asset class

Facility asset class	**Number of defaults**		**Number of loans**		**Exposure [in bn EUR]**	
Small/Medium enterprises	67.40%	106,640	68.32%	210,521	17.54%	131,682
Large corporates	14.14%	22,371	16.12%	49,679	47.88%	359,533
Banks and financial companies	2.03%	3,216	1.88%	5,804	12.91%	96,941
Ship finance	0.62%	987	0.57%	1,758	2.64%	19,852
Aircraft finance	0.25%	388	0.29%	892	0.94%	7,063
Real estate finance	10.79%	17,068	8.76%	27,001	10.96%	82,293
Project finance	0.38%	598	0.41%	1,254	2.41%	18,099
Commodities finance	0.29%	457	0.27%	842	1.32%	9,900
Sovereigns, central banks	0.09%	147	0.09%	284	1.71%	12,833
Public services	0.15%	234	0.12%	366	0.33%	2,515
Private banking	3.87%	6,124	3.16%	9,747	1.36%	10,248
Total	100%	158,230	100%	308,148	100%	750,959

Panel B: Share of defaulted loans by region

Region	**Defaults**
Europe	58%
North America	33%
Asia	5%
Africa	2%
Middle East	<1%
Oceania	<1%

2.2 LGD Estimation

Financial institutions typically develop statistical models based on historical default data to predict the LGDs of their borrowers. For this purpose, they have a large number of methods at their disposal, ranging from traditional linear regression to advanced methods of machine learning. Because of the wide range, the most common methods are used in the empirical analyses in Chapters 3-5. The methods are categorized as either traditional or advanced methods. The methods (summarized in Table 2.2) and the main references for each are presented below.

Linear regression (LR) is used as the first traditional method because it usually serves as a reference method in other LGD studies. For instance, the linear regression has been implemented in a comparative context by Loterman et al. (2012) and Krüger and Rösch (2017). From a