Machine learning-based variable selection for clustered credit risk modeling

Marc Gürtler*, Marvin Zöllner**

University of Braunschweig – Institute of Technology, Department of Finance

Abstract

Several studies have demonstrated the high prediction accuracy of clustered credit risk modeling. In clustered modeling, borrowers are segmented based on their similarities through cluster analysis, and a separate predictive 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, optimal variable selection for clustering is a major challenge, particularly for high-dimensional data. In the present study, we propose a machine learning-based variable selection method. Formally, the most influential risk drivers identified from the machine learning model using Shapley values are used as clustering variables. Thus, the information of the explanatory variables crucial for the prediction of the dependent variable is already processed during data segmentation, making each individual predictive model more effective. Through a comparative analysis using two real-world credit default datasets, we show that clustered modeling with cluster analysis optimized by machine learning-based variable selection leads to the highest prediction accuracy.

Keywords: Credit Risk, Forecasting, Clustering, Machine Learning, Global Credit Data

JEL Classification: C38, C45, C52, C53, G21

*Corresponding author. University of Braunschweig – Institute of Technology, Department of Finance, Abt-Jerusalem-Straße 7, 38106 Braunschweig, Germany, Phone: +49 531 391 2895, Email: marc.guertler@tu-bs.de.

**University of Braunschweig – Institute of Technology, Department of Finance, Abt-Jerusalem-Straße 7, 38106 Braunschweig, Germany, Phone: +49 531 391 2894, Email: marvin.zoellner@tu-bs.de.

Declarations of interest: none. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The provision of data by Global Credit Data (*GCD*) is gratefully acknowledged.

1 Introduction

Credit risk modeling is an important risk management task in financial institutions. In this context, banks develop a statistical model based on historical default data to predict borrowers' credit risk and derive targeted risk management strategies, such as adjusting lending policies. However, given the varying nature of borrowers, a single statistical model may not be sufficient to capture the risk characteristics of various individuals. Clustered modeling can be used to overcome this problem. In this approach, borrowers are clustered based on their similarities through cluster analysis and for each resulting cluster a separate predictive model is developed. Because the borrowers in each cluster have similar risk characteristics, the models can be individually developed and fitted to each cluster, resulting in higher predictive accuracy (see, for instance, Bakoben et al. (2020)). The effectiveness of clustered modeling depends on the quality of borrower segmentation, which in turn is primarily influenced by the choice of explanatory variables used for clustering. It has long been known that not every variable is useful for cluster structure detection, and the inclusion of irrelevant variables may impair the ability of clustering procedures to effectively detect meaningful structures (e.g., De Soete et al. (1985), Milligan (1989), and Green et al. (1990)). Precisely, the use of inappropriate variables that possess no discriminative information for clustering may result in overlapping, indistinguishable, and uninformative clusters (cf. Fop & Murphy (2018)), which negatively affects the predictive performance in separate modeling. Therefore, the selection of optimal variables used in cluster analysis is particularly challenging, especially for high-dimensional data.[1] The difficulty also arises from the fact that high-dimensional data can be meaningfully clustered in a variety of ways. More specifically, it is not necessary to identify the variables that lead to the best clustering but rather those that enable the best prediction of the dependent variable in separate modeling.

To address the variable selection challenge in clustering, we propose an intelligent variable selection process optimized using machine learning.[2] More formally, our approach is to calibrate a high-precision machine learning model and then use the Shapley values of Shapley (1953) as an importance measure to determine the risk drivers that most strongly affect the estimations. In the next step, only those variables whose Shapley values exceed an importance threshold are used in the cluster analysis. In this way, we use exactly those variables in clustering that are most important in the machine learning model and thus make the greatest contribution to the explanation of the dependent variable. Finally, we obtain an optimal set of variables that contains the essential information for predicting the dependent variable. Based on this variable set, cluster analysis leads to highly informative clusters, which, in turn, improve the performance of predictive models in separate modeling. Hereafter we refer this procedure to as "optimized clustered approach." From the set of machine learning algorithms, we choose gradient-boosted trees by Friedman (2001) and random forest by Breiman (2001) (in the robustness check) as intelligent variable selection models, because several studies have already shown their superiority in credit risk modeling (e.g., Lessmann et al. (2015) and Gürtler & Zöllner (2023a))

This study focuses on modeling loss given default (*LGD*), which is one of the main drivers of credit risk associated with credit products. High predictive accuracy is essential in LGD modeling for several reasons. First, by predicting LGD accurately, banks can identify high-risk borrowers and adjust lending policies to minimize the risk of loss from borrower defaults. Second, accurate LGD prediction

[1] Even for data with moderate or low dimensionality, reducing the set of variables employed in the clustering process can be beneficial (Fowlkes et al. (1988)).

[2] Machine learning algorithms are also increasingly being used in other areas of finance. For instance, see Apel et al. (2022).

is crucial for loan pricing. Incorrect LGD predictions can lead to incorrect pricing, resulting in greater losses or lower profitability. Third, regulatory authorities require banks to implement robust credit risk management practices. Accurate LGD predictions are essential to comply with these requirements and demonstrate that banks have adequate capital to cover potential losses. In summary, banks use LGD to make risk-based decisions and accurate predictions can result in significant competitive advantage, whereas weak predictions can lead to adverse selection.

To investigate the effectiveness of our optimized clustered approach, we conduct an intensive benchmark study. Specifically, we apply the optimized clustered approach and competing approaches (including a standard (non-clustered) approach and clustered approaches with baseline techniques for variable selection in clustering) to a dataset of defaulted loans from US enterprises. We find that clustered approaches generally lead to higher predictive accuracies than the standard (non-clustered) approach. Most importantly, we show that our optimized clustered approach considerably outperforms competing clustered approaches. In this context, we find that clustering based on the three most important risk drivers for LGD leads to the optimal clustering, which significantly improves the out-of-sample performance of the predictive models in separate modeling. Moreover, the optimized clustered approach creates economically meaningful and comprehensible clusters, as required by the regulators. These results are robust to various indicators of predictive accuracy and are confirmed by a robustness check. In the robustness check, we use a European credit portfolio, that is, European empirical data with different loan characteristics compared to the US data, to ensure that the superiority of the optimized clustered approach does not depend on the choice of a particular dataset.

This study contributes to the literature on clustered credit risk modeling by proposing an optimal variable selection method for clustering using machine learning. In literature, the challenge of variable selection has been addressed in three ways. The simplest selection is no selection; that is, all available variables are often used for clustering (e.g., Harris (2015) and Caruso et al. (2021)). However, this approach may be suitable for low-dimensional data. Nevertheless, considering all the variables in many cases unnecessarily increases the complexity of the clustering process. In addition, some variables may not have any relevant information for predicting the dependent variable, and consequently, should not be used for clustering. Rather, they can adversely affect the quality of clustering by increasing the likelihood of overlapping clusters, thereby reducing the accuracy of predictive models in separate modeling. Second, the literature proposes the use of baseline techniques for variable selection, with most studies using principal component analysis (*PCA*) (e.g., Yoshino & 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, which are referred to as components. Thus, variables are automatically selected for clustering but at the cost of a lower understanding of meaning. However, regulators generally require explainability in credit risk modeling[3], which actually limits the practical applicability of PCA as a variable selection technique. The third way to select variables for clustered modeling is to use linear regression with stepwise variable elimination (e.g., Yuan et al. (2022)). In this procedure, the variables to be used for clustering are selected from a set of candidate variables using a linear regression model through a series of automated steps. Specifically, at each step, the candidate variables are iteratively used in linear regression and in-sample evaluated, typically using the t-statistics for the coefficients of the considered variables. Finally, variables with the

[3] In credit risk modeling, explainability is generally required in Article 179(1)(a) of the Capital Requirements Regulation (CRR) (see European Banking Authority (2013)).

highest statistical significance in the linear regression model are used for clustering. However, a fundamental problem with this procedure is that through iterative testing, some explanatory variables that actually have causal effects on the dependent variable may not be statistically significant, while irrelevant variables may be significant by chance (cf. Smith (2018)). As a result, an implausible and inefficient set of variables may be identified, which negatively affects the quality of clustering and thus reduces the accuracy of predictive models in separate modeling.

Against this background, we present an approach that enables intelligent variable selection for clustered modeling and has several advantages. First, using machine learning algorithms, the optimal variables for clustering are automatically and effectively identified, considerably reducing the risk of creating uninformative clusters. Second, unlike other variable selection techniques, our approach uses original variables for clustering, ensuring the transparency and interpretability of the cluster results, as recommended by regulators. Third, the approach is agnostic; that is, it is applicable to any model for variable selection, and there are no restrictions on the use of specific data (i.e., in terms of size or dimensionality), making the approach suitable for a wide range of applications.

The remainder of this paper is organized as follows. In Section 2, we introduce our optimized clustered approach based on arbitrary machine learning algorithms for variable selection. Section 3 presents the empirical data and settings used for the comparative analyses and describes the competitive modeling approaches. In Section 4, we compare the out-of-sample performance of all the modeling approaches using various evaluation criteria. Section 5 presents a robustness check. Finally, Section 6 concludes the paper.

2 Optimized clustered model

In this section, we describe the optimized clustered approach, which is schematically illustrated in Figure 1. This approach consists of four steps, described in detail below.

In the first step, we divide the entire dataset into a subsample for training (in-sample calibration) and a subsample for testing (out-of-sample prediction), as is common in LGD studies (e.g., Hartmann-Wendels et al. (2014) and Hurlin et al. (2018))

Next, we calibrate the machine learning model (in our case, gradient-boosted trees or random forest) using a set M of all available explanatory variables z_1, z_2, ..., $z_{|M|}$ based on the training dataset. Calibration of machine learning models involves many parameters (such as the number of regression trees in gradient-boosted trees) that must be determined. We optimize and choose these parameters by a process known as hyperparameter tuning.[4] In this context, we determine the parameter values using a five-fold cross-validation (e.g., Nazemi et al. (2017) and Hurlin et al. (2018)) and a grid search algorithm: The in-sample dataset is divided into five subsets, four parts of which serve as training data and the remaining part as test data. This procedure is repeated five times using different test datasets. In this process, the grid search algorithm trains the machine learning model based on all possible hyperparameter settings, where the hyperparameters are selected from a predefined hyperparameter set. Finally, the parameter values with the highest estimation accuracy (for example, the smallest mean squared er-

[4] For a detailed description of the tree-based models and parameters to be determined, see, e.g., (Hastie et al., 2017, 305 et seqq.).

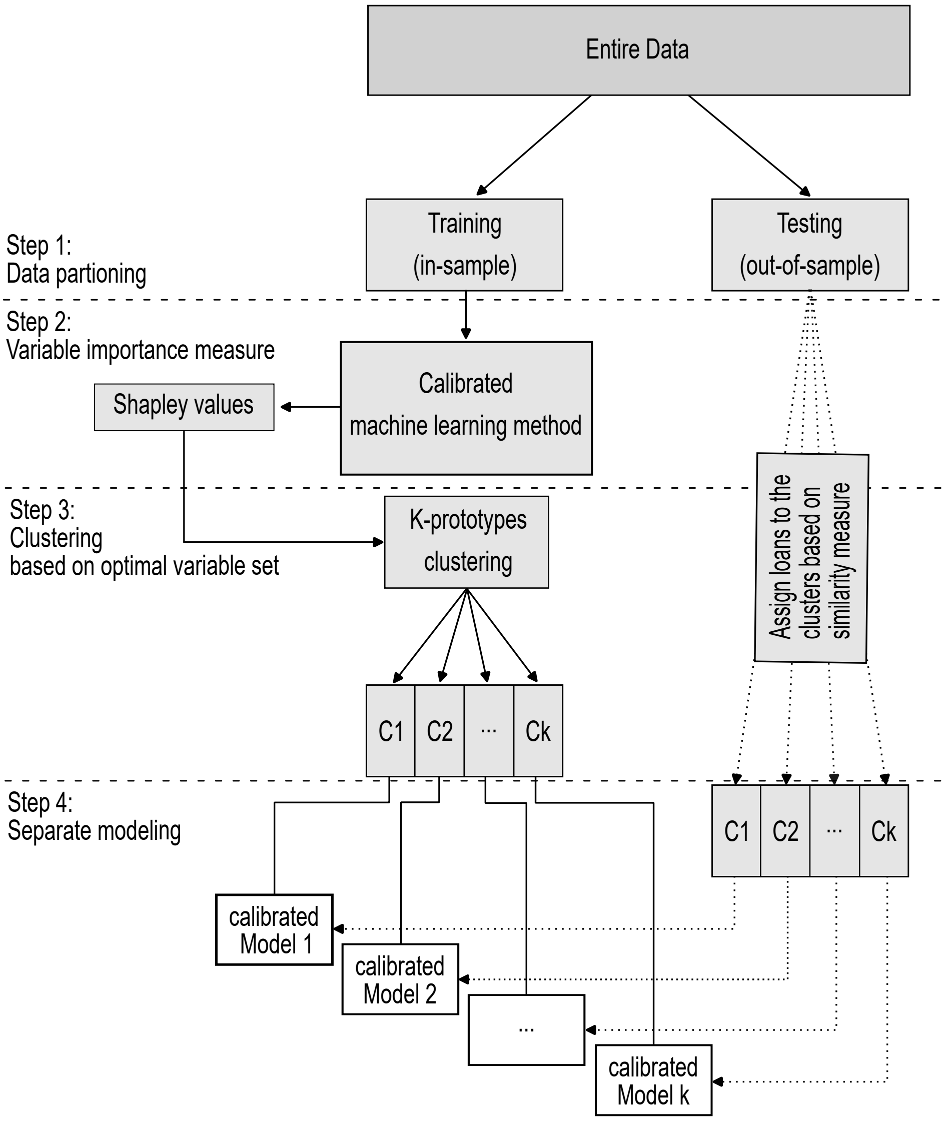

***Figure 1*. Schematic diagram of the optimized clustered approach.**

ror on the test set) are selected.[5] After calibrating the machine learning model, we calculate the mean Shapley value, originally introduced by Shapley (1953), for each explanatory variable. The concept of Shapley values comes from cooperative game theory and is used to identify the most important variables in the machine learning model and select the variables for clustering. Simply put, the Shapley value represents for a concrete data observation the contribution of a certain explanatory variable in estimating the dependent variable when added to arbitrary variable sets. To measure the global importance of each variable, we average the variable contributions of all individual observations. Because the determination of the Shapley value is highly computationally intensive for high-dimensional data, we use the Tree SHAP algorithm by Lundberg & Lee (2017), which approximates Shapley values with relatively low computational intensity.[6]

[5] The hyperparameters necessary for gradient-boosted trees and random forest, the corresponding sets of considered parameter values, and the final choice of hyperparameter values used in the empirical analysis are listed in Table OA.3.

[6] For more details on variable importance with Shapley values, see Gürtler & Zöllner (2023b).

The third step involves the optimized clustering process. In this context, we first standardize[7] the training data and perform clustering using the k-prototypes algorithm by Huang (1998), which is an improvement of the k-means and k-mode algorithms to handle clustering with mixed data types. The algorithm aims to partition the dataset into k clusters (with randomly selected k observations as initial cluster centers) such that the distances between observations, characterized by a given set of explanatory variables, are minimized within a cluster, and the distances between different clusters are maximized. The similarity measure is a combination of the Euclidean distance for numeric variables and a simple matching approach for categorical variables. The contribution of this study is to improve clustering such that the resulting clusters enable the highest prediction accuracy in separate modeling. This is achieved by performing optimal variable selection, which is the basis for clustering. For this purpose, we optimize the clustering process (Step 3), which can be described as follows:

- Step 3.1: Sort explanatory variables in descending order of their mean Shapley values. Without loss of generality, let the resulting rank order be given by $z_1, z_2, ..., z_M$.

- Step 3.2: Loop the number k of clusters from k_{min} to k_{max}[8] and iteratively partition the data based on an increasing set of the most important variables in each loop. That is, the k-prototypes clustering is first based on z_1, then on z_1 and z_2, then on z_1, z_2, and z_3, et cetera.

- Step 3.3: Calculate the silhouette value[9] by Rousseeuw (1987) for every combination (k; (z_1, ..., z_i)) of the number of clusters and variable sets, to evaluate the quality of the resulting clusters.

- Step 3.4: Repeat steps 3.1 - 3.3 (e.g., 10,000 times) with the initial cluster centers changed. This is because clustering algorithms are sensitive to the initial cluster centers.

- Step 3.5: Average the silhouette values (hereafter referred to as "global silhouette value") of each combination of number of clusters and variable set. The combination with the highest global silhouette value leads to optimal clustering, thereby determining the final number k^* of clusters and the optimal variable set (z_1, ..., z_i^*).

- Step 3.6: Choose the k^* optimal clusters $c_1, c_2, ..., c_{k^*}$, based on the optimal variable set.

In the final step, we perform separate modeling. This involves back-standardizing the data and calibrating a separate predictive model based on all available explanatory variables z_1, z_2, ..., $z_{|M|}$ for each resulting cluster, resulting in k^* different models being used for prediction. For prediction, individual out-of-sample loans are assigned to the respective (cluster) subsamples based on the same similarity measure used in the k-prototype algorithm.[10]

For reasons of clarity and comprehensibility, we specify some important terms. In the following we consider the entire "clustered modeling approach" (Figure 1) as combination of a "clustering model" and a "prediction model.". The clustering model consists of a clustering method (k-prototypes) and a variable selection method (in this case, the measurement of the importance of the variables of a machine

[7] To standardize the data is recommended; otherwise, the range of values of each variable may serve as a weight in determining the clustering of data, which is usually undesirable.

[8] In the empirical analysis we set $k_{min} = 2$ and $k_{max} = 10$.

[9] This coefficient is calculated as $(b - a)/max(a, b)$ using the mean within-cluster distance (a) and the mean next nearest-cluster distance (b). The highest (and most preferred) value is 1 and the lowest value is -1.

[10] Additionally, we tested other measures to validate the consistency of the cluster results and used different methods to initialize the cluster centers. However, these changes did not affect our results.

learning method based on Shapley values). The prediction model consists of a prediction method (e.g., random forest or linear regression) and model parameter choice (calibration); that is, each prediction method is calibrated based on k clusters, resulting in k (cluster-specific) prediction models.

3 Empirical framework

To demonstrate the effectiveness of our optimized clustered approach, we conduct an intensive benchmark study. For this purpose, we use the Global Credit Data[11] database, which contains detailed information on the credit defaults of 55 banks, including many systemically important banks. It is internationally recognized as the standard for collecting LGD data because it is officially approved to be in line with regulatory guidelines. In the following section, we first introduce the data and provide descriptive statistics. Next, we describe the competitive modeling approaches used in the benchmark study and explain the procedure and measures for comparing their out-of-sample performance.

3.1 Data

We use a dataset of resolved defaulted loans from small- and medium-sized enterprises ($SMEs$) and large corporations (LCs) in the US. We use these two asset classes because they are categorized as general corporate exposures under the regulatory guidelines. To calculate LGD, we use workout recovery rates, which are given as the difference between all discounted post-default incoming cash flows (F^+) and all discounted post-default costs (C^-), divided by the exposure at default (EAD). That is,

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

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 using the three-month LIBOR of the respective default date.

Below, we briefly describe the restrictions we apply to the raw dataset, which includes 10,516 defaulted loans, to ensure consistency and plausibility. All restrictions are based on recommendations from the LGD literature (cf. European Banking Authority (2016), Betz et al. (2018), and Gürtler & Zöllner (2023a)). First, 572 observations are excluded due to time span restrictions. Specifically, we restrict the sample to all defaults since 2000 to ensure a consistent default definition of Basel II and exclude defaults after 2019. This upper bound is selected for two reasons. First, workout processes of recent defaults are not necessarily completed. Additionally, in the subsample of recently defaulted loans (with uncompleted workout processes), short workout periods are obviously overrepresented. As loans with shorter workout periods tend to be associated with lower LGDs, this subsample can lead to a sample selection bias, which may result in unreliable estimation results. Second, in the Global Credit Data database, the default amounts range from zero (e.g., for uncalled contingent facilities) to several hundred million euros. To meet the materiality threshold required by regulators, we remove loans with an EAD of less than \$500, which leads to the exclusion of 211 observations. Third, we exclude 52 observations by correcting for minor input errors. That is, we eliminate loans with an abnormally low or high LGD; that

[11] See `https://www.globalcreditdata.org`.

is, smaller than −100% and higher than 200%, respectively. Finally, loans with incomplete observations are excluded, thus we remove 224 observations. Overall, a dataset of 9,457 loans remains.

Table 1
Descriptive statistics.
Note. This table shows the means and quantiles of the loan characteristics, macroeconomic factors, and empirical LGDs (in %) for various loan categories.

Variable	**Level**	**Quantiles**					**Mean**	**Obs.**
		0.05	**0.25**	**0.50**	**0.75**	**0.95**		
$\text{LGD}_{overall}$		−5.21	−0.05	5.01	44.32	99.78	24.31	9457
log(EAD)		9.23	11.41	12.86	14.49	16.75	12.92	9457
Number of collaterals		0.00	0.00	1.00	2.00	5.00	1.39	9457
Number of guarantors		0.00	0.00	0.00	1.00	2.00	0.42	9457
LGD conditional to loan categories:								
Guarantee indicator	No guarantee	−4.94	−0.02	5.99	47.10	100.00	25.21	6607
	Guarantee	−5.83	−0.19	3.32	39.78	98.70	22.23	2850
Collateral indicator	No collateral	−2.35	3.77	14.64	53.28	100.00	30.97	1076
	Collateral	−5.36	−0.17	3.84	42.71	99.21	23.46	8381
Facility type	Medium term	−5.28	−0.03	6.99	44.39	99.78	25.33	8768
	Short term	−4.38	−0.47	4.45	42.23	99.54	24.05	689
Seniority type	Pari-passu	−1.69	2.69	14.25	47.79	99.42	28.19	3113
	Super senior	−5.99	−1.01	1.36	40.64	99.76	22.14	6218
	Non senior	−0.94	0.38	14.84	79.95	100.00	35.80	126
Facility asset class	Small/Medium	−5.86	−0.69	2.74	42.41	100.00	23.26	6829
	Large	−1.62	1.74	13.04	47.39	97.29	27.06	2628
Syndication indicator	No syndication	−5.41	−0.15	4.24	43.60	99.93	23.90	8751
	Syndication	−0.24	4.09	16.89	50.20	94.84	29.47	706
Lender limit	No Limit	−1.60	1.99	11.72	49.39	100.00	28.14	3564
	Limit	−6.12	−1.15	1.31	39.91	99.49	22.00	5893
Borrower type	Public	−1.70	1.70	14.64	49.70	96.97	28.02	1202
	SPV	0.41	8.85	9.63	18.65	53.19	18.29	57
	Private	−5.47	−0.22	3.95	43.07	99.93	23.79	8204
Industry type								
Finance, insurance, real estate	(FIRE)	−6.23	−1.07	2.66	29.89	96.84	18.83	1397
Agriculture, forestry, fishing, hunting	(AFFH)	−7.05	−0.19	0.55	6.41	92.99	12.87	190
Mining	(MIN)	−0.59	−0.14	1.40	30.39	93.17	19.70	263
Construction	(CON)	−5.45	−0.74	2.58	45.06	98.71	23.10	1324
Manufacturing	(MAN)	−3.68	0.24	8.24	45.15	97.68	25.17	1190
Transp., commu.,elec., gas, sani. serv.	(TCEGS)	−5.18	0.24	11.78	51.39	96.48	27.40	778
Wholesale and retail trade	(WRT)	−4.88	0.23	11.43	52.61	100.00	29.14	871
Services	(SERV)	−5.88	−0.57	3.83	49.64	100.00	25.72	2340
Other	(Other)	−2.72	1.09	9.57	40.23	100.00	25.88	1104
S&P 500 (rel. change)		−38.49	−12.91	6.16	14.51	30.40	1.55	9457
3-month LIBOR (abs. spread in p. p.)		0.23	0.29	0.60	2.20	5.37	1.55	9457
Term spread (abs. spread in p. p.)		−0.18	1.62	2.43	3.15	3.55	2.21	9457
TED spread (abs. spread in p. p.)		0.15	0.20	0.30	0.54	1.44	0.49	9457
10-year bond yield (abs. spread in p. p.)		1.72	2.52	3.40	4.10	5.16	3.37	9457
Cboe volatility index (abs. spread in p. p.)		12.09	15.89	20.70	26.35	44.14	22.94	9457
GDP growth rate (annual %)		−3.29	0.50	1.72	2.61	3.87	1.21	9457
Inflation rate (annual %)		−0.36	1.26	1.64	3.16	3.84	1.90	9457
Unemployment rate (annual %)		4.40	5.10	6.80	9.00	9.90	7.04	9457
Consumer confidence index		57.30	69.50	76.40	84.50	95.70	77.64	9457
Producer price index		131.20	167.90	181.90	196.90	204.00	176.85	9457
Consumer price index		−0.47	−0.10	0.17	0.44	0.84	0.14	9457

Table 1 presents the descriptive statistics. Specifically, we report the means and several quantiles of the metric variables. For each level of categorical variables, we show the means and category-specific quantiles of the respective LGD as well as the number of observations per group. The table provides an indicator of the plausibility of the dataset. For example, the existence of guarantees or securities reduces

LGDs. Conversely, non-senior and medium-term loans lead to higher LGDs. We also distinguish between other loan categories, such as facility asset classes, syndication, lender limits, types of borrowers, and firms' industry affiliation.

In addition to the loan-specific characteristics, we also consider various macroeconomics control variables to improve the prediction of the LGD, as suggested in the literature.[12] Stock exchange performances are identified as general LGD risk drivers, for instance, by Qi & Zhao (2011) and Chava et al. (2011). To consider the overall real and financial environment in the US, we use the relative year-on-year growth of the S&P 500, the absolute spread of the three-month and 10-years treasury rates, the absolute term and TED spread, and the Cboe volatility index. We also use the annual percentage growth rate of gross domestic product to measure the market value of all final goods and services produced in the considered period (cf. Yao et al. (2015)). Moreover, we consider other popular macroeconomic variables, such as the inflation rate, unemployment rate, consumer confidence index, producer price index, and consumer price index. A detailed description of the variables is provided in Table OA.2 in the online appendix. Specific macroeconomic information corresponds to the default time of each loan. All macroeconomic data is provided by Refinitiv Eikon[13].

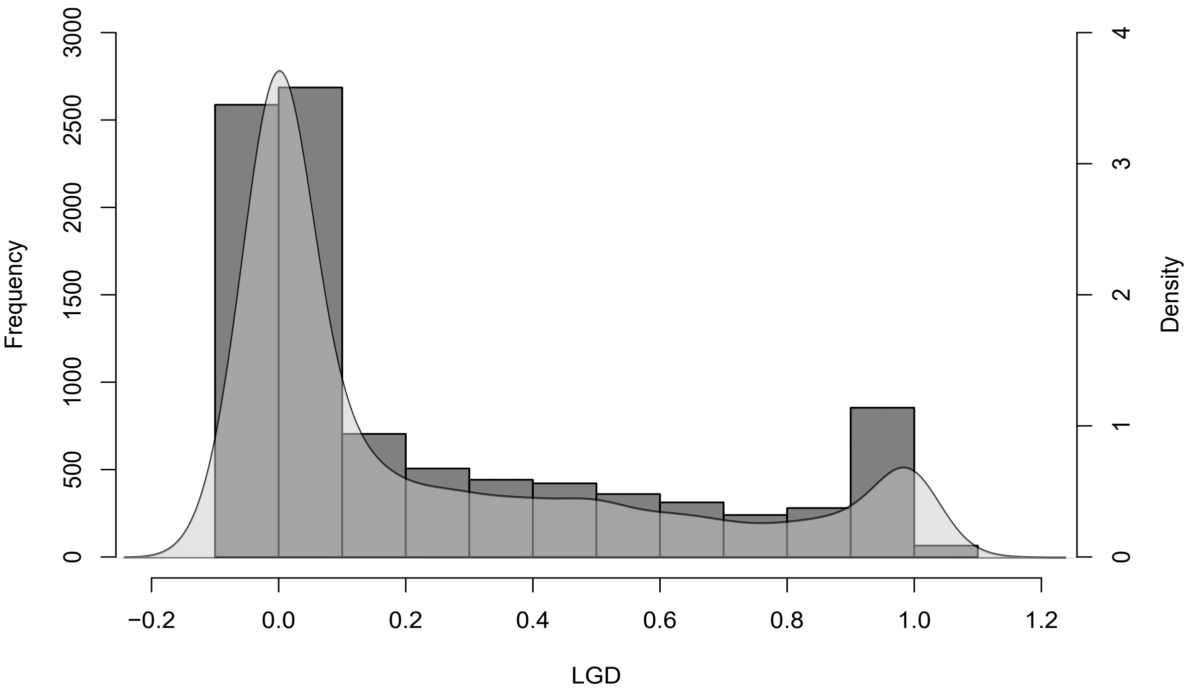

***Figure 2*. LGD frequency and approximated density distribution.**

Figure 2 shows the LGD frequency and approximated density distribution. Most LGDs represent (nearly) total losses or recoveries, yielding strong bimodality and skewness of the distribution. These properties reinforce our decision to use gradient-boosted trees as an intelligent variable selection model for clustering because they have been shown by Gürtler & Zöllner (2023a) to adequately capture bimodality and skewness.

[12] See, for instance, Nazemi et al. (2017). New technical standards emphasize the importance of using economic factors (European Banking Authority (2017)).

[13] See `https://www.refinitiv.com` for further information.

3.2 Competitive modeling approaches

To fairly evaluate the effectiveness of our optimized clustered approach, we must compare its predictive performance with that of a standard (non-clustered) approach and various clustered approaches using different clustering models.[14] In each modeling approach, different models (calibrated prediction methods) are used for prediction. This leads to an investigation of different combinations of clustering and prediction models with the aim of identifying the best combination with the highest prediction accuracy. In the following section, we first describe the competing modeling approaches and then introduce the prediction models considered.

To demonstrate the superiority of the clustered modeling approaches, we first apply a standard modeling approach. Precisely, after splitting the entire dataset into training and test data, one predictive model is calibrated based on the (non-clustered) training data and applied to the test data for out-of-sample prediction. However, this standard approach to credit risk modeling should be improved by clustered modeling (cf. Bakoben et al. (2020)).

Additionally, we use clustered modeling approaches that follow the same scheme as in Figure 1 (see Section 2) but differ in the selection of variables for clustering. In total, we consider five different clustering models based on recommendations from the literature[15] First, we consider the most basic clustering model that uses all available explanatory variables z_1, z_2, ..., $z_{|M|}$ to cluster the training data (e.g., Caruso et al. (2021)). Second, we use a clustering model with a silhouette decomposition algorithm for variable selection. The algorithm partitions the training data based on the explanatory variables that provide the best clustering without considering the variables' ability to predict the dependent variable in separate modeling (cf. Dessureault & Massicotte (2021)). Third, we use a clustering model that employs linear regression with the k-best algorithm to select the variables for clustering. A linear regression model that includes all the variables z_1, z_2, ..., $z_{|M|}$ is first calibrated based on the training data. Subsequently, the variables are selected using the k-best algorithm, which scores the variables of the linear regression model using the F-test, and then all but the k highest scoring variables are removed with respect to their p-values (see, for instance, Jain & Verma (2022)). The variables remaining in the linear regression model are used for clustering. Fourth, we use a clustering model similar to the previous one; however, instead of using the k-best algorithm, we use a stepwise elimination algorithm to select the variables for clustering. This algorithm is a hybrid version of forward selection and backward elimination. It begins with a linear regression model that contains no variables, and the variables are then selected as in forward selection; that is, the variables that contribute the most to the model fit in terms of the p-value are iteratively added to the model. After each step, the variables are checked for elimination according to backward elimination; that is, the variables with the smallest contribution to the model fit are eliminated. The idea behind this is that, with the addition of new variables, the variables already considered in the model could become redundant and should therefore be removed. (e.g., Loterman et al. (2012)). The variables remaining in the linear regression model are used for clustering. Fifth, we use a clustering model that employs factor analysis for mixed data ($FAMD$) to select variables for clustering. FAMD generalize PCA to categorical and numerical data and is used to reduce the dimensionality of training data while preserving as much as possible of the information contained in the original data. As previously mentioned, this aim is achieved

[14] We note that all competing clustered approaches use the k-prototypes algorithm as the clustering method but differ in the method of variable selection for clustering.

[15] We note that in each clustering model, the number k of clusters are iteratively varied between $k = 2$ and $k = 10$ and this is repeated 10,000 times with changing initial cluster centers.

by creating new variables, referred to as components, as linear combinations or mixtures of the initial variables (cf. Le et al. (2021)). To determine the number of components to be used for clustering, the training data is iteratively partitioned based on an increasing number of components, and the set with the best clustering (i.e., highest global silhouette value) is selected.

In the following, we briefly introduce the models (calibrated prediction methods) used in competing approaches for prediction. Because there is a wide range of predictive models used in the LGD literature, we apply the most established models. The optimal hyperparameter values for the machine learning models are determined in the same manner as described in Section 2. [16]

First, we use linear regression because it is typically used as a reference model in other LGD studies. For instance, the linear regression has been implemented in a comparative context byLoterman et al. (2012) and Krüger & Rösch (2017). However, from a statistical perspective, linear regression has certain restrictions that may render it unsuitable for LGD estimation. Therefore, we also include machine learning models that address these restrictions.

As the first machine learning models, we use various tree-based models because they allow nonparametric representations of the relationships between the dependent and explanatory variables. The most basic model in this class is the regression tree, which was popularized by Breiman (1984). Briefly, it recursively splits the data into groups and uses the group averages of the dependent variable as its mean prediction. This model has been applied to LGD estimation by, for instance, Matuszyk et al. (2010) and Hurlin et al. (2018). In addition, we use random forest by Breiman (2001) and gradient-boosted trees by Friedman (2001) as extensions of the simple regression tree. The former is a bootstrap aggregation model of decorrelated regression trees built independently using random subsets of variables and trained on different parts of the same training set. In contrast, in boosting, trees are built sequentially, and each tree is constructed based on the residual errors made by the previous tree, leading to a nonrandom model that generates fewer prediction errors as more trees are added. The use of random forest and gradient-boosted trees for LGD prediction is proposed by, for example, Bastos (2014) and Tanoue & Yamashita (2019).

In addition to tree-based models, we also consider a multilayer perceptron model proposed by, for instance, Bishop (1995) and support vector regression introduced by Vapnik (1995). The former is a fully connected class of feedforward artificial neural network that consists of several highly interconnected processing elements that process information by their dynamic state response to external inputs. To calculate the network, we use a resilient backpropagation algorithm that guarantees an approximation of the estimation value through iterative model updates. Support vector regression extends the linear regression by considering nonlinear relationships in the coefficients. The main idea is to map the data into a higher-dimensional space using a mapping function (in our case, the radial-basis function kernel) before performing linear regression.[17]

3.3 Empirical setup

In this section, we describe the empirical setup used to compare the predictive performance of competitive modeling approaches. The dataset is divided into a subsample for training and a subsample for

[16] The hyperparameters necessary for each predictive model, the corresponding sets of considered parameter values, and the final choice of hyperparameter values are available upon request.

[17] We refer interested readers to Hastie et al. (2017) for a more detailed description of neural network and support vector regression.

testing. For the training dataset, we use data from 2000 to 2013, which correspond to approximately 70% of the entire dataset. Subsequently, we randomly draw 10,000 times a subsample from the remaining data from 2014 to 2019. Each step consists of 500 defaulted loans, which is approximately the average number of defaults per year for the entire dataset. In this process, we apply the calibrated models (of each modeling approach) to each testing subsample and evaluate their predictive accuracy out-of-sample (and out-of-time). To measure the predictive performance, we use four popular criteria: mean squared error (MSE), mean absolute error (MAE), median absolute error ($MedAE$), and coefficient of determination (R^2), which are defined as follows:

$$MSE := \frac{1}{n}\sum_{i=1}^{n}(LGD_i - \widehat{LGD}_{i,m})^2 \tag{2}$$

$$MAE := \frac{1}{n}\sum_{i=1}^{n}|LGD_i - \widehat{LGD}_{i,m}| \tag{3}$$

$$MedAE := median(|LGD_1 - \widehat{LGD}_{1,m}|, ..., |LGD_n - \widehat{LGD}_{n,m}|) \tag{4}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(LGD_i - \widehat{LGD}_{i,m})^2}{\sum_{i=1}^{n}(LGD_i - \overline{LGD})^2} \tag{5}$$

where n corresponds to the number of observations in the respective dataset, LGD_i denotes the true LGD value of the i^{th} credit, $\widehat{LGD}_{i,m}$ denotes the corresponding LGD estimation using method m, and $\overline{LGD}$ corresponds to the arithmetic mean of the true LGD values. Because the $MedAE$ is more resistant to outliers, we use it in combination with the MAE. A high absolute difference between the MAE and $MedAE$ indicates that there are outliers among the estimation errors. Finally, the mean of each criterion calculated over all 10,000 steps denotes the predictive accuracy of the respective modeling approach.

Based on the out-of-sample criteria, the modeling approaches can be ranked from worst to best. To exclude the possibility that some superiority may have occurred by chance, we complement the standard performance measures with the model confidence set (MCS) procedure of Hansen et al. (2011). This procedure involves statistical tests that allow a set of modeling approaches to be identified that are "superior" with a given probability (i.e., confidence level α)[18]. Thereby, sequential hypothesis testing on the null hypothesis of equal predictive ability (EPA) between competing modeling approaches is utilized. The MCS procedure is as follows. We start with an initial set of approaches of dimension d. In the next step, we test the EPA null hypothesis. If this hypothesis is rejected, the approach with the lowest performance is removed from the set of potentially superior approaches, and the algorithm repeats this step with the reduced set of approaches. If the null hypothesis is not rejected, the algorithm terminates, and the remaining d^* approaches define the superior set $\widehat{D}^*_{1-\alpha}$. The superior set does not have to be single-element, that is, besides $d^* = 1$, $1 < d^* \leq d$ is possible. We use the MCS procedure to individually compare competitive approaches based on the test MSEs and test MAEs, respectively.

[18] For the confidence level, we set $\alpha = 0.1\%$.

4 Empirical results

As already mentioned, we use gradient-boosted trees as the variable selection method within the optimized clustering model because they are particularly suitable for capturing the properties of the LGD distribution of US data (bimodality and skewness). In this section, we first present the results of determining the variable importance in gradient-boosted trees using Shapley values. Next, the clustering results of all the competitive clustering models are presented, and those of our optimized clustering model is described in more detail. Finally, we state and evaluate the results of the comparative out-of-sample analyses.

4.1 Variable importance measure

After calibrating the gradient-boosted trees within the optimized clustering model, we identify the most influential variables for estimating the LGD. Figure 3 shows the ranking of the global variables importance, resulting from the determination of the mean absolute contributions of all variables. The results confirm the findings from the literature on key LGD risk drivers (e.g., Dermine & de Carvalho (2006), Grunert & Weber (2009), Krüger & Rösch (2017), and Betz et al. (2018)) and can be summarized as follows:

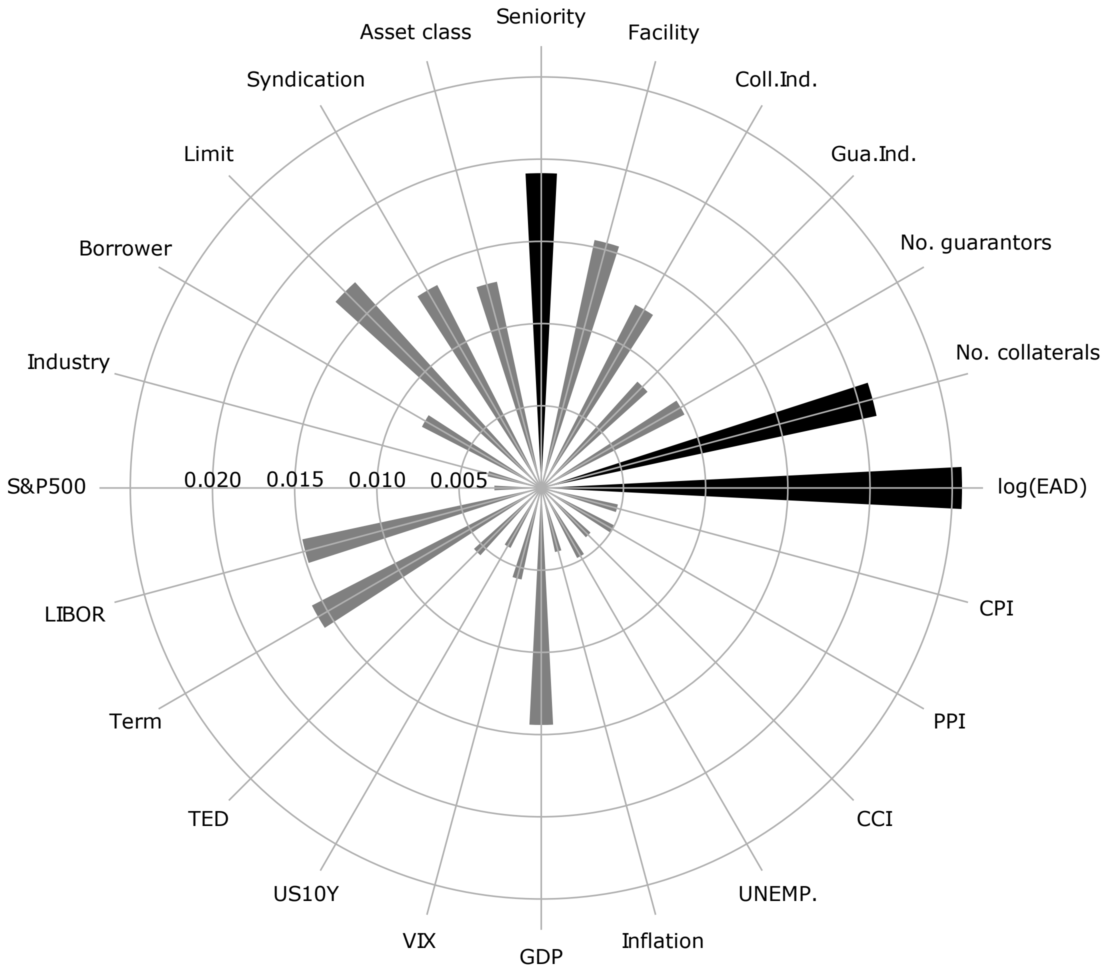

***Figure 3*. Variable importance measure in gradient-boosted trees used within the optimized clustering model.**

First, log(EAD) most strongly affects the estimates of the gradient-boosted trees. Second, collateral-related variables, such as the number of collaterals, seniority, or limit have the next largest affect on the estimates of gradient-boosted trees. Third, collateral is more relevant than guarantees for estimating the LGD. Fourth, the company's industry affiliation does not seem to have a relevant effect on LGD. Fifth, while the macroeconomic variables GDP, term, and LIBOR are considered relatively important, the other macroeconomic variables seem to play only a minor role in estimating LGD. In summary, we conclude that both macroeconomic and loan-specific variables matter; however, EAD and collateral-related variables are especially crucial for LGD estimation in gradient-boosted trees.

4.2 Cluster analysis

Figure 4 shows the clustering results of the optimized clustering model. The key findings are as follows: First, if the number i^* of important variables used for clustering is too large ($i^* \geq 10$), the quality of the clustering (in terms of global silhouette value) decreases significantly. Second, a similar result is obtained for the number k of clusters. For most number i^* of important variables, an increasing number k of clusters leads to a reduction in the global silhouette value, indicating that the more complex the clustering process (in terms of i^* and k), the worse the final clustering result. Third, the crucial result is that using the $i^* = 3$ most important variables (log(EAD), no. collaterals, and seniority) in clustering leads to the best result for three clusters, with a global silhouette value of approximately 0.50.

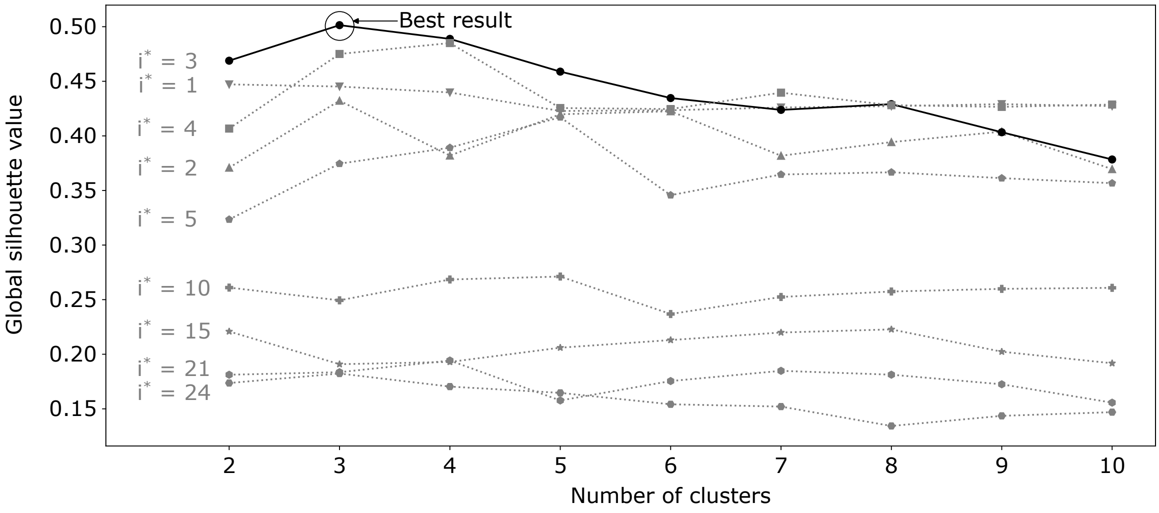

***Figure 4*. Clustering results of the optimized clustering model.**
Note. The number of important variables used for clustering is indicated by i^*.

Figure 5 compares the clustering results of all the competing clustering models. It becomes clear that the model which uses $i^* = 24$ variables for clustering, that is, without specific variable selection, leads to lower quality of the resulting clusters compared to the clustering models with variable selection. More specifically, clustering using all available variables leads to the worst overall result, with a global silhouette value of less than 0.2. In addition, for each clustering model, the quality of the resulting clusters is strongly related to the number k of clusters. For instance, for the model with the k-best variable selection, the global silhouette value is reduced from initially around 0.35 for three clusters to approximately 0.25 for ten clusters. Unsurprisingly, the clustering model that uses the silhouette decomposition algorithm for variable selection has the highest global silhouette value, with a value greater than 0.6 for two clusters. However, this algorithm aims to generate the best homogeneously separated clusters without considering the relevance of individual variables in the prediction of LGD. Although this leads to optimal clustering with respect to all variables, it neglects the fact that only a few variables are relevant for prediction, and clustering should, therefore, only take place on the basis of these variables. This also explains why the set of variables used in this clustering model (asset class, industry, and borrower) is completely different from those used in the optimized cluster model. In summary, all other clustering models using the baseline methods for variable selection achieve a lower cluster quality than the optimized clustering model.

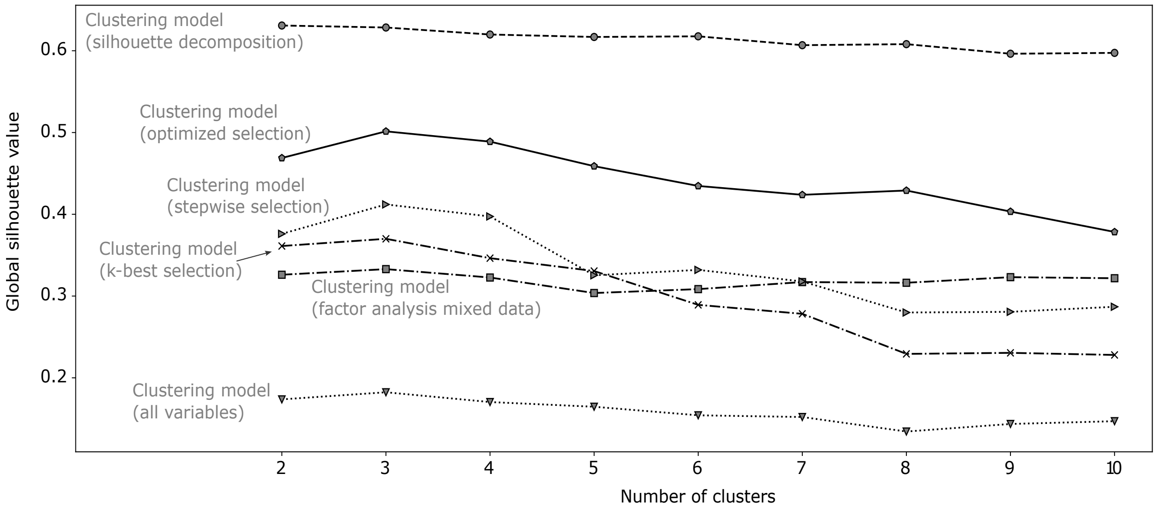

***Figure 5*. Comparison of the clustering results of the competing clustering models.**
Note. The variable selection methods used in the clustering models are indicated in parentheses. Additionally, the variables used in clustering are specified for each model.
Clustering model (all variables): All variables
Clustering model (silhouette decomposition): Asset class, Industry, Borrower
Clustering model (factor analysis mixed data): Four components
Clustering model (k-best selection): No. collaterals, No. guarantors, Coll.Ind., Facility, Borrower, Industry, PPI
Clustering model (stepwise selection): log(EAD), No. guarantors, Coll.Ind., Limit, Borrower, Industry, VIX, UNEMP.
Clustering model (optimized selection): log(EAD), No. collaterals, Seniority

To characterize the resulting clusters of the optimized clustering model, the mean values of the numeric variables and modal values of the categorical variables are listed for each cluster in Table 2. The results can be summarized as follows. The first cluster comprises predominantly medium-term loans from small/medium-sized enterprises characterized by a low average LGD and log(EAD), a high average number of collaterals, and a super senior status. The second cluster included loans of the same asset class, maturity, and seniority, with significantly increased average LGD and log(EAD) and a reduced number of collaterals. The third cluster consists mainly of medium-term loans with pari-passu status from large corporations, which are characterized by a particularly high average LGD and log(EAD), as well as a low number of collaterals. In summary, the loans in the training dataset are clustered into three segments characterized by low, medium, and high average LGD. This segmentation is plausible because the three modal values (close to zero, 0.5, and close to one) are already evident in the LGD distribution of the entire dataset (Figure 2). Therefore, we can confirm that the optimized clustering model leads to economically meaningful and comprehensible clusters, as required by regulators.

Table 2
Interpretation of the resulting clusters of the optimized clustering model.
Note. This table shows the means of the numerical variables and modal values of the categorical variables for each cluster.

Variable	Cluster 1	Cluster 2	Cluster 3
LGD	0.06	0.41	0.78
log(EAD)	10.57	12.27	14.64
No. collaterals	8.30	3.21	0.62
No. guarantors	0.81	0.44	0.14
Gua.Ind.	Yes	No	No
Coll.Ind.	Yes	Yes	Yes
Facility	Medium	Medium	Medium
Seniority	Super senior	Super senior	Pari-Passu
Asset class	Small/Medium	Small/Medium	Large
Syndication	No	No	No
Limit	Yes	Yes	No
Borrower	Private	Private	Private
Industry	SERV	SERV	SERV
S&P500	7.44	3.38	−1.85
LIBOR	1.00	1.32	2.05
Term	2.44	2.27	2.04
TED	0.48	0.50	0.47
US10Y	2.97	3.16	3.81
VIX	22.68	22.76	23.12
GDP	1.21	1.15	1.38
Inflation	1.83	1.97	2.07
UNEMP.	6.4	7.26	7.83
CCI	73.216	75.91	81.3
PPI	189.06	182.53	165.17
CPI	0.11	0.13	0.17
Total observations	3382	1453	1784

4.3 Comparative analysis

In this subsection, we present the results of the comparative out-of-sample analysis. Table 3 lists the performances of competing modeling approaches. As mentioned earlier, each modeling approach is a combination of a clustering model and prediction model.

First, we find that each clustered approach performs better than the standard approach without clustering. For example, the MSE and MAE of the prediction models within the standard approach vary between 0.1099 (gradient-boosted trees) and 0.1245 (linear regression) and 0.2745 and 0.3032, respectively. In contrast, the MSEs and MAEs are substantially reduced, even for the prediction models within the simplest clustered approach (with the worst clustering quality) using all variables in clustering, varying between 0.1042 and 0.1202 and between 0.2686 and 0.2957, respectively. Therefore, we confirm the results in the literature, indicating the higher accuracy of clustered approaches compared to the standard (non-clustered) modeling approach.

Second, we can conclude that the clustered approach with the “best” (i.e., highest global silhouette value) partitioning of the training dataset using variable selection based on silhouette decomposition does not have the best prediction performance at the same time. Although it performs better than the non-clustered approach and the simplest clustered approach, it is considerably outperformed by other clustered approaches using clustering models with baseline methods for variable selection. For example, the coefficient of determination of the prediction models in the clustered approach using stepwise variable selection is approximately 3%-6% higher. Moreover, the prediction models within the clustered approach using variable selection based on silhouette decomposition seem to be more influenced by outliers, as shown by the larger differences between the MAE and MedAE.

Table 3

Results of the comparative out-of-sample analysis.

Note. Each modeling approach is a combination of a clustering model and prediction model. The variable selection methods used in the clustering models are indicated in parentheses. The final assessment of the prediction models is based on the average of each criterion calculated for all 10,000 samples. The models marked with (*) are identified in the MCS procedure as the superior set within the same modeling approach. Models marked with (‡) are identified as the superior set across all competitive modeling approaches.

Clustering model	Prediction model	MSE	MAE	|MAE - MedAE|	R^2
No clustering	Linear regression	0.1245	0.3032	0.0650	0.1217
	Regression tree	0.1230	0.2899	0.0313*	0.1499
	Random forest	0.1122*	0.2750*	0.0341	0.1867
	Gradient-boosted trees	0.1099*	0.2745*	0.0312*	0.1883*
	Support vector regression	0.1162	0.2784	0.0367	0.1782
	Multilayer perceptron	0.1193	0.2863	0.0567	0.1723
Clustering model (all variables)	Linear regression	0.1202	0.2957	0.0633	0.1423
	Regression tree	0.1195	0.2823	0.0212*	0.1727
	Random forest	0.1077	0.2711	0.0290	0.2123
	Gradient-boosted trees	0.1042*	0.2686*	0.0244	0.2292*
	Support vector regression	0.1113	0.2735	0.0312	0.1989
	Multilayer perceptron	0.1092	0.2728	0.0255	0.2134
Clustering model (silhouette decomposition)	Linear regression	0.1197	0.2876	0.0538	0.1689
	Regression tree	0.1186	0.2782	0.0159	0.1791
	Random forest	0.1058	0.2678	0.0246	0.2286
	Gradient-boosted trees	0.1038*	0.2622*	0.0125*	0.2311*
	Support vector regression	0.1092	0.2692	0.0255	0.2121
	Multilayer perceptron	0.1094	0.2732	0.0262	0.2110
Clustering model (k-best selection)	Linear regression	0.1132	0.2759	0.0387	0.1816
	Regression tree	0.1121	0.2655	0.0154	0.1843
	Random forest	0.0984	0.2648	0.0124*	0.2424*
	Gradient-boosted trees	0.0971*	0.2614*	0.0115*	0.2463*
	Support vector regression	0.1019	0.2655	0.0186	0.2388
	Multilayer perceptron	0.0998	0.2649	0.0168	0.2373
Clustering model (stepwise selection)	Linear regression	0.1102	0.2727	0.0328	0.2048
	Regression tree	0.1094	0.2625	0.0109	0.2072
	Random forest	0.0962*	0.2615*	0.0093*	0.2492
	Gradient-boosted trees	0.0949*	0.2582*	0.0089*	0.2555*
	Support vector regression	0.0979	0.2613	0.0131	0.2477
	Multilayer perceptron	0.0974	0.2631	0.0143	0.2482
Clustering model (factor analysis mixed data)	Linear regression	0.1114	0.2734	0.0345	0.1982
	Regression tree	0.1102	0.2631	0.0120	0.2036
	Random forest	0.0961*	0.2611*	0.0080*	0.2487
	Gradient-boosted trees	0.0952*	0.2589*	0.0088*	0.2544*
	Support vector regression	0.0994	0.2623	0.0145	0.2381
	Multilayer perceptron	0.0972	0.2625	0.0132	0.2458
Clustering model (optimized selection)	Linear regression	0.1085	0.2661	0.0219	0.2187
	Regression tree	0.1004	0.2615	0.0150	0.2389
	Random forest	0.0921	0.2553	0.0020*	0.2736
	Gradient-boosted trees	0.0887*‡	0.2501*‡	0.0017*‡	0.2798*‡
	Support vector regression	0.0933	0.2587	0.0074	0.2715
	Multilayer perceptron	0.0914	0.2538	0.0031	0.2744

Third, using variable selection for clustering leads to a better performance of the prediction models. For example, while the MSEs of the prediction models within the clustered approach using all variables in clustering vary between 0.1042 and 0.1202, they vary between 0.0952 and 0.1114 for prediction models within the clustered approach with factor analysis, and between 0.0949 and 0.1102 for prediction models within the clustered approach with stepwise selection. This result is consistent for all evaluation criteria.

Fourth, in each modeling approach (i.e., regardless of the choice of the clustering model), the gradient-boosted trees are superior to the other models in predicting the LGD for each evaluation criterion. To exclude the possibility that this superiority occurred by chance, we perform the MCS procedure across all prediction models within each modeling approach. We find that the gradient-boosted trees are always identified as the significantly superior set of models when compared based on each evaluation criterion. This result reinforces our decision to use gradient-boosted trees as an intelligent variable selection method in our optimized clustered approach.

Fifth, the most important result of the comparative analysis is that the optimized clustered approach has a higher prediction accuracy than the other clustered approaches, regardless of the specific choice of the predictive model used in the separate modeling. For example, for gradient-boosted trees, we observe significant improvements in the MSE, MAE, and R^2 of approximately 20%, 9%, and 48%, respectively. Owing to the small differences between MAEs and MedAEs, the prediction models within the optimized clustered approach are also robust against outliers in the prediction errors. Interestingly, even the performance of the simple linear regression is remarkably improved in the optimized clustered approach.

To determine the overall best combination of clustering model and prediction model, we perform the MCS procedure across all possible combinations. We find that the optimized clustering model, together with gradient-boosted trees, are identified as the superior combination and lead to the highest prediction accuracy. Overall, we confirm the superiority of our clustered modeling approach optimized using machine learning. Additionally, we find that not the best but optimal clustering, in the sense of using variables in clustering that contain relevant information for predicting LGD, leads to the best out-of-sample predictive performance.

5 Robustness check

To ensure that the superiority of the optimized clustered approach does not depend on the choice of specific data, we use a European credit portfolio with other loan characteristics in this robustness check. Using 3,137 defaulted loans by small, medium, and large enterprises, provided by Global Credit Data, we rerun our comparative analysis. Precisely, based on the same empirical setup, the prediction models used within the competitive modeling approaches are re-calibrated, optimal hyperparameter values are re-determined, and cluster analysis, out-of-sample model comparisons, and significance tests are re-performed. The restrictions applied to the data are the same as those applied to the US data. The descriptive statistics and LGD distribution of the European dataset are shown in Table OA.1 and Figure OA.2 in the online appendix. We observe a (nearly) symmetric bimodal LGD distribution with total losses and total recoveries being equally likely. Because Gürtler & Zöllner (2023a) have recently shown that random forest provides the best out-of-sample predictions for data with this distribution type, we use it as an intelligent variable selection method in the optimized clustered approach.

Figure 6 shows the ranking of the global variables importance in the random forest for the European data. Similar to the US data, we find that log(EAD) and collateral-related variables are crucial for LGD estimation, with asset class and limit becoming more important. The crucial difference, however, is that for the European data random forest assigns little importance to all macroeconomic variables.

The comparison of the clustering results of the competitive clustering models is shown in Figure

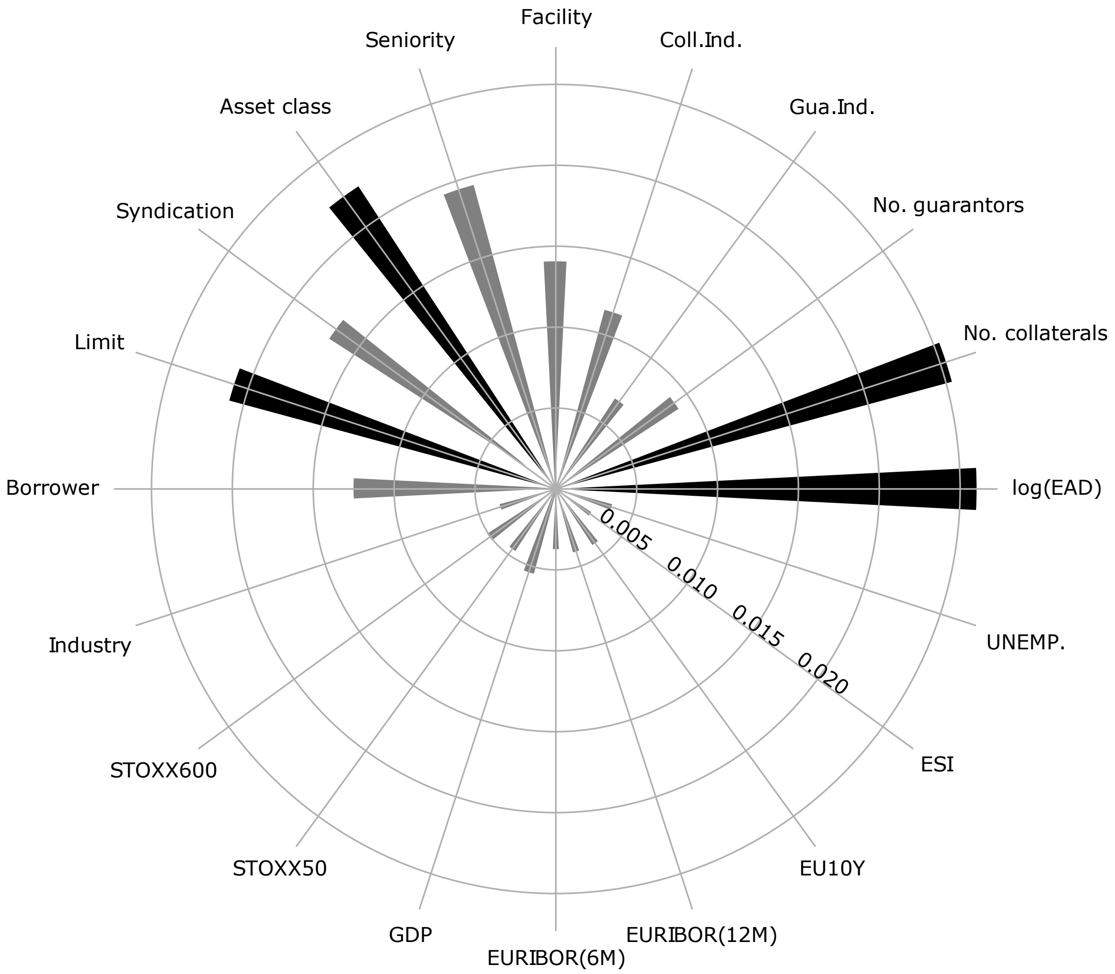

***Figure 6*. EU Data: Variable importance measure in random forest used within the optimized clustering model.**

7.[19] The conclusions we draw for the US data can also be confirmed for European data. First, the more complex the clustering process (in terms of i^* and k), the lower the quality of the resulting clusters. Second, all clustered approaches using clustering models with baseline methods for variable selection achieve a lower cluster quality than the optimized clustered approach. Third, using three clusters and the $i^* = 4$ (instead of $i^* = 3$ for the US data) most important variables (log(EAD), no. collaterals, asset class, and limit) leads to the best cluster result for the optimized clustered approach.

Table 4 presents the mean values of the numerical variables and modal values of the categorical variables for each cluster created using the optimized clustered approach. Similar to the analysis of the US data, the loans in the training dataset are clustered into three segments characterized by low, medium, and high average LGD, corresponding to the three identifiable modal values in the LGD distribution of the entire data (cf. Figure OA.2).

The out-of-sample performances of the competing modeling approaches (i.e., a combination of clustering and prediction models) are shown in Table 5. We confirm the superiority of the clustered approaches over the standard approach without clustering. In addition, the clustered approach with

[19] The clustering results of the optimized clustered approach and clustered approach with factor analysis are shown in detail in Figures OA.3 and OA.4 in the online appendix.

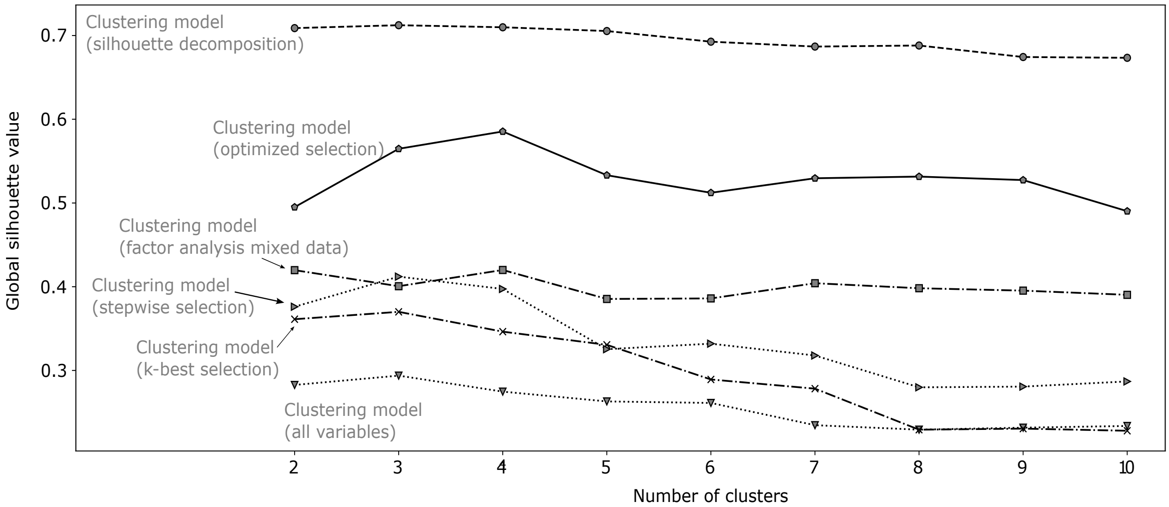

***Figure 7*. EU data: Comparison of the clustering results of the competing clustering models.**

Note. The variable selection method used in the clustering model is indicated in parentheses. Additionally, the variables used in clustering are specified for each model.
Clustering model (all variables): All variables
Clustering model (silhouette decomposition): Facility, Limit, Seniority
Clustering model (factor analysis mixed data): Three components
Clustering model (k-best selection): Industry, No. guarantors, STOXX600, Seniority, Borrower, UNEMP., Syndication
Clustering model (stepwise selection): No. collaterals, Coll.Ind., Facility, Borrower, Industry, EURIBOR(6M), UNEMP.
Clustering model (optimized selection): Log(EAD), No. collaterals, Asset class, Limit

Table 4
EU data: Interpretation of the resulting clusters of the optimized clustering model.

Note. This table shows the means of the numerical variables and modal values of the categorical variables for each cluster.

Variable	Cluster 1	Cluster 2	Cluster 3
LGD	0.11	0.45	0.82
log(EAD)	11.94	12.33	13.62
No. collaterals	1.5	1.01	0.53
No. guarantors	0.21	0.13	0.02
Gua.Ind.	No	No	No
Coll.Ind.	Yes	Yes	No
Facility	Medium	Medium	Medium
Seniority	Super senior	Super senior	Pari-Passu
Asset class	Small/Medium	Small/Medium	Large
Syndication	No	No	No
Limit	Yes	No	No
Borrower	Private	Private	Private
Industry	WRT	WRT	AFFH
STOXX600	−0.54	−0.58	−8.78
STOXX50	−3.00	−3.16	−12.14
EURIBOR(6M)	2.01	2.20	2.50
EURIBOR(12M)	2.21	2.35	2.61
EU10Y	4.23	3.88	4.00
GDP	0.02	0.01	0.01
UNEMP.	9.30	9.4	9.62
ESI	94.36	93.28	95.02
Total observations	700	947	548

Table 5

EU data: Results of the comparative out-of-sample analysis.

Note. Each modeling approach is a combination of a clustering model and prediction model. The variable selection methods used in the clustering models are indicated in parentheses. The final assessment of the prediction models is based on the average of each criterion calculated for all 10,000 samples. The models marked with (⋆) are identified in the MCS procedure as the superior set within the same modeling approach. Models marked with (‡) are identified as the superior set across all competitive modeling approaches.

Clustering model	Prediction model	MSE	MAE	|MAE - MedAE|	R^2
No clustering	Linear regression	0.1239	0.3112	0.0679	0.1221
	Regression tree	0.1211	0.2954	0.0442	0.1512
	Random forest	0.1089⋆	0.2858⋆	0.0402⋆	0.1926⋆
	Gradient-boosted trees	0.1152	0.2913	0.0316⋆	0.1844
	Support vector regression	0.1151	0.2908	0.0411	0.1828
	Multilayer perceptron	0.1197	0.2931	0.0423	0.1711
Clustering model (all variables)	Linear regression	0.1212	0.2946	0.0465	0.1506
	Regression tree	0.1189	0.2873	0.0332	0.1765
	Random forest	0.1063⋆	0.2771⋆	0.0180⋆	0.2244⋆
	Gradient-boosted trees	0.1112	0.2807	0.0209	0.1992
	Support vector regression	0.1112	0.2795	0.0217	0.1995
	Multilayer perceptron	0.1134	0.2841	0.0320	0.1892
Clustering model (silhouette decomposition)	Linear regression	0.1207	0.2942	0.0463	0.1473
	Regression tree	0.1186	0.2870	0.0321	0.1785
	Random forest	0.1062⋆	0.2766⋆	0.0165⋆	0.2249⋆
	Gradient-boosted trees	0.1102	0.2801	0.0194	0.2049
	Support vector regression	0.1097	0.2796	0.0214	0.2110
	Multilayer perceptron	0.1112	0.2832	0.0302	0.2098
Clustering model (k-best selection)	Linear regression	0.1185	0.2931	0.0402	0.1762
	Regression tree	0.1172	0.2844	0.0278	0.1793
	Random forest	0.1048⋆	0.2710⋆	0.0116⋆	0.2282⋆
	Gradient-boosted trees	0.1051⋆	0.2753	0.0165	0.2272⋆
	Support vector regression	0.1061	0.2767	0.0189	0.2388
	Multilayer perceptron	0.1055⋆	0.2749	0.0167	0.2268
Clustering model (stepwise selection)	Linear regression	0.1182	0.2915	0.0444	0.1771
	Regression tree	0.1155	0.2826	0.0288	0.1805
	Random forest	0.1029⋆	0.2687⋆	0.0105⋆	0.2303⋆
	Gradient-boosted trees	0.1049	0.2789	0.0145	0.2280
	Support vector regression	0.1032⋆	0.2727⋆	0.0085⋆	0.2294⋆
	Multilayer perceptron	0.1089	0.2812	0.0184	0.1928
Clustering model (factor analysis mixed data)	Linear regression	0.1153	0.2883	0.0372	0.1811
	Regression tree	0.1121	0.2797	0.0199	0.1873
	Random forest	0.0982⋆	0.2653⋆	0.0041⋆	0.2421⋆
	Gradient-boosted trees	0.1021	0.2744	0.0131	0.2381
	Support vector regression	0.0999	0.2697	0.0088	0.2404
	Multilayer perceptron	0.1064	0.2789	0.0191	0.2362
Clustering model (optimized selection)	Linear regression	0.1093	0.2793	0.0184	0.2112
	Regression tree	0.1065	0.2783	0.0165	0.2374
	Random forest	0.0892⋆‡	0.2592⋆‡	0.0011⋆‡	0.2755⋆‡
	Gradient-boosted trees	0.0931	0.2633	0.0026	0.2694
	Support vector regression	0.0922	0.2627	0.0019	0.2711
	Multilayer perceptron	0.0987	0.2692	0.0058	0.2564

the "best" partitioning of the training dataset using variable selection based on silhouette decomposition does not have the highest prediction accuracy. In addition, for European data, modeling approaches using variable selection for clustering have better performances than the simplest clustered approach using all variables in clustering. Moreover, each prediction model within the optimized clustered approach show higher predictive accuracy than the prediction models within the other clustered approaches. Overall, the two basic conclusions of the analyses – the superiority of the optimized clustered approach and the need for optimal clustering rather than the best clustering – can also be drawn for European data.

6 Conclusion

Banks typically use statistical models to predict borrowers' credit risks. However, many academic studies have shown that a single (non-clustered) model may not be sufficient to capture the risk characteristics of various individual borrowers, and therefore propose the use of clustered modeling. In this approach, borrowers are segmented based on their similarities through cluster analysis, and a separate predictive model is developed for each cluster, resulting in high predictive performance.

The main challenge with clustered approaches is selecting the optimal variables used in the cluster analysis, especially for high-dimensional data. An incorrect choice can result in overlapping, indistinguishable, and uninformative clusters, which negatively affect the predictive performance in the separate modeling. Moreover, high-dimensional 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 dependent variable in separate modeling.

Against this background, we propose a clustered approach with an intelligent variable selection process for clustering optimized using machine learning models. As part of this approach, we automatically and effectively identify variables that contain relevant information for predicting credit risk and use these variables in a cluster analysis, which considerably reduces the risk of creating uninformative clusters. Moreover, a particular advantage of our approach is that it is independent of the machine learning model used for variable selection, and thus has a high degree of flexibility in its application.

The superiority of our optimized clustered approach is investigated through an empirical analysis using two real-life LGD datasets. We demonstrate that the optimized clustered approach outperforms non-clustered modeling and clustered approaches using baseline methods for variable selection. This conclusion is robust to several indicators of predictive accuracy. Moreover, we show that our optimized clustered approach creates economically meaningful and comprehensible clusters as required by regulators and provides interesting insights into the influence of explanatory variables on LGD. In particular, we find that exposure at default and collateral-related variables are crucial in LGD modeling.

References

Apel, M., Blix Grimaldi, M., & Hull, I. (2022). How much information do monetary policy committees disclose? evidence from the fomc's minutes and transcripts. *Journal of Money, Credit and Banking*, *54*(5), 1459–1490. doi: 10.1111/jmcb.12885

Bakoben, M., Bellotti, T., & Adams, N. (2020). Identification of credit risk based on cluster analysis of account behaviours. *Journal of the Operational Research Society*, *71*(5), 775–783. doi: 10.1080/01605682.2019.1582586

Bastos, J. A. (2014). Ensemble predictions of recovery rates. *Journal of Financial Services Research*, *46*(2), 177–193. doi: 10.1007/s10693-013-0165-3

Betz, J., Kellner, R., & Rösch, D. (2018). Systematic effects among loss given defaults and their implications on downturn estimation. *European Journal of Operational Research*, *271*(3), 1113–1144. doi: 10.1016/j.ejor.2018.05.059

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford: Oxford University Press and Clarendon Press.

Breiman, L. (1984). *Classification and regression trees*. New York: Chapman & Hall/CRC.

Breiman, L. (2001). Random Forest. *Machine Learning*, *45*(1), 5–32. doi: 10.1023/A:1010933404324

Caruso, G., Gattone, S. A., Fortuna, F., & Di Battista, T. (2021). Cluster analysis for mixed data: An application to credit risk evaluation. *Socio-Economic Planning Sciences*, *73*, 100850. doi: 10.1016/j.seps.2020.100850

Chava, S., Stefanescu, C., & Turnbull, S. (2011). Modeling the loss distribution. *Management Science*, *57*(7), 1267–1287. doi: 10.1287/mnsc.1110.1345

Dermine, J., & de Carvalho, C. N. (2006). Bank loan losses-given-default: a case study. *Journal of Banking & Finance*, *30*(4), 1219–1243. doi: 10.1016/j.jbankfin.2005.05.005

De Soete, G., DeSarbo, W. S., & Carroll, J. D. (1985). Optimal variable weighting for hierarchical clustering: An alternating least-squares algorithm. *Journal of Classification*, *2*(1), 173–192. doi: 10.1007/BF01908074

Dessureault, J.-S., & Massicotte, D. (2021). Feature selection or extraction decision process for clustering using pca and frsd. *Working paper*. doi: 10.48550/arXiv.2111.10492

European Banking Authority. (2013). Article 179 of the capital requirements regulation (crr): Regulation (eu) no 575/2013 of the european parliament and of the council of 26 june 2013 on prudential requirements for credit institutions and investment firms and amending regulation (eu) no 648/2012. Retrieved from https://www.eba.europa.eu/regulation-and-policy/single-rulebook/interactive-single-rulebook/1746

European Banking Authority. (2016). Guidelines on PD estimation, LGD estimation and the treatment of defaulted exposures. *Consultation Paper*.

European Banking Authority. (2017). Impact assessment for the GLs on PD, LGD and the treatment of defaulted exposures based on the IRB survey results. *EBA Report on IRB modelling practices*.

Fop, M., & Murphy, T. B. (2018). Variable selection methods for model-based clustering. *Statistics Surveys*, *12*. doi: 10.1214/18-SS119

Fowlkes, E. B., Gnanadesikan, R., & Kettenring, J. R. (1988). Variable selection in clustering. *Journal of Classification*, *5*(2), 205–228. doi: 10.1007/BF01897164

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232. doi: 10.1214/aos/1013203451

Green, P. E., Kim, J., & Carmone, F. J. (1990). A preliminary study of optimal variable weighting in k-means clustering. *Journal of Classification*, *7*(2), 271–285. doi: 10.1007/BF01908720

Grunert, J., & Weber, M. (2009). Recovery rates of commercial lending: empirical evidence for German companies. *Journal of Banking & Finance*, *33*(3), 505–513. doi: 10.1016/j.jbankfin.2008.09.002

Gürtler, M., & Zöllner, M. (2023a). Heterogeneities among credit risk parameter distributions: the modality defines the best estimation method. *OR Spectrum*, *45*(1), 251–287. doi: 10.1007/s00291-022-00689-6

Gürtler, M., & Zöllner, M. (2023b). Tuning white box model with black box models: Transparency in credit risk modeling. *SSRN Electronic Journal*. doi: 10.2139/ssrn.4433967

Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The model confidence set. *Econometrica*, *79*(2), 453–497. doi: 10.3982/ECTA5771

Harris, T. (2015). Credit scoring using the clustered support vector machine. *Expert Systems with Applications*, *42*(2), 741–750. doi: 10.1016/j.eswa.2014.08.029

Hartmann-Wendels, T., Miller, P., & Töws, E. (2014). Loss given default for leasing: parametric and nonparametric estimations. *Journal of Banking & Finance*, *40*, 364–375. doi: 10.1016/j.jbankfin.2013.12.006

Hastie, T., Tibshirani, R., & Friedman, J. H. (2017). *The elements of statistical learning: data mining, inference, and prediction* (Second edition, corrected at 12th printing 2017 ed.). New York, NY: Springer.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values.

Hurlin, C., Leymarie, J., & Patin, A. (2018). Loss functions for Loss Given Default model comparison. *European Journal of Operational Research*, *268*(1), 348–360. doi: 10.1016/j.ejor.2018.01.020

Jain, A., & Verma, D. (2022). Making credit underwriting process more accurate using ml. In *2022 international conference on advances in computing, communication and materials (icaccm)* (pp. 1–4). IEEE. doi: 10.1109/ICACCM56405.2022.10009117

Krüger, S., & Rösch, D. (2017). Downturn LGD modeling using quantile regression. *Journal of Banking & Finance*, *79*, 42–56. doi: 10.1016/j.jbankfin.2017.03.001

Le, R., Ku, H., & Jun, D. (2021). Sequence-based clustering applied to long-term credit risk assessment. *Expert Systems with Applications*, *165*, 113940. doi: 10.1016/j.eswa.2020.113940

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, *247*(1), 124–136. doi: 10.1016/j.ejor.2015.05.030

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. doi: 10.1016/j.ijforecast.2011.01.006

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon et al. (Eds.), *Advances in neural information processing systems* (Vol. 30, pp. 4765–4774). Curran Associates, Inc. Retrieved from `https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf`

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. doi: 10.1057/jors.2009.67

Milligan, G. W. (1989). A validation study of a variable weighting algorithm for cluster analysis. *Journal of Classification*, *6*(1), 53–71. doi: 10.1007/BF01908588

Nazemi, A., Fatemi Pour, F., Heidenreich, K., & Fabozzi, F. J. (2017). Fuzzy decision fusion approach for loss-given-default modeling. *European Journal of Operational Research*, *262*(2), 780–791. doi: 10.1016/j.ejor.2017.04.008

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

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. doi: 10.1016/0377-0427(87)90125-7

Shapley, L. S. (1953). A value for n-person games. In H. Kuhn & A. W. Tucker (Eds.), *Contributions to the theory of games, ii.* Princeton University Press.

Smith, G. (2018). Step away from stepwise. *Journal of Big Data*, *5*(1). doi: 10.1186/s40537-018-0143-6

Tanoue, Y., & Yamashita, S. (2019). Loss given default estimation: a two-stage model with classification tree-based boosting and support vector logistic regression. *Journal of Risk*, *21*(4), 19–37. doi: 10.21314/JOR.2019.405

Vapnik, V. N. (1995). *The nature of statistical learning theory*. New York, NY: Springer. doi: 10.1007/978-1-4757-2440-0

Yao, X., Crook, J., & Andreeva, G. (2015). Support vector regression for loss given default modelling. *European Journal of Operational Research*, *240*(2), 528–538. doi: 10.1016/j.ejor.2014.06.043

Yoshino, N., & Taghizadeh-Hesary, F. (2019). A comprehensive method for credit risk assessment of small and medium-sized enterprises based on asian data. In N. Yoshino & F. Taghizadeh-Hesary (Eds.), *Unlocking sme finance in asia* (pp. 55–71). First Edition. | New York : Routledge, 2019. | Series: Routledge studies in development economics: Routledge. doi: 10.4324/9780429401060-3

Yuan, K., Chi, G., Zhou, Y., & Yin, H. (2022). A novel two-stage hybrid default prediction model with k-means clustering and support vector domain description. *Research in International Business and Finance*, *59*, 101536. doi: 10.1016/j.ribaf.2021.101536

Online appendix / Supplementary materials

Figures

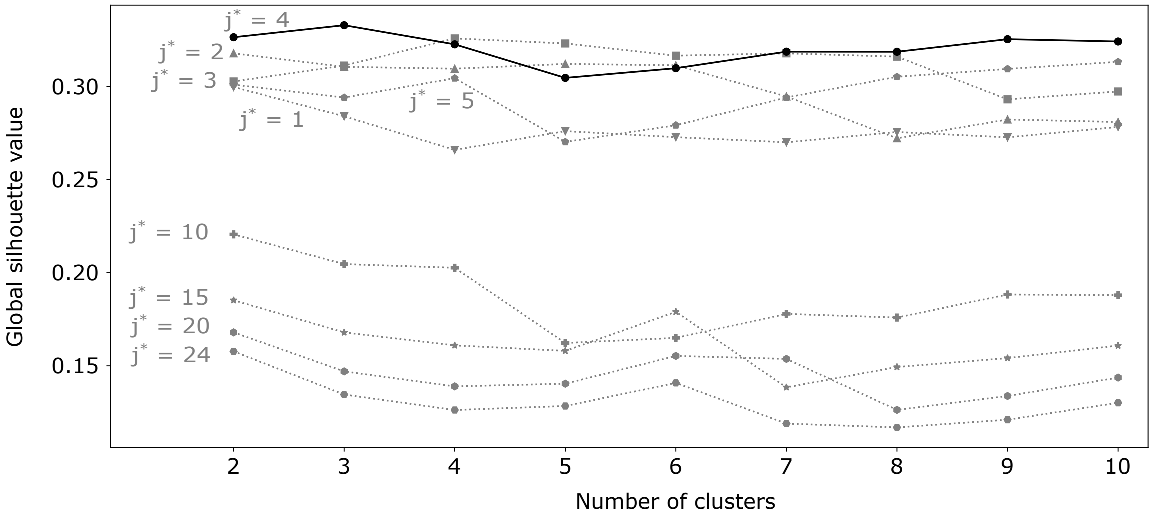

***Figure OA.1*. US data: Clustering results of the clustering model using factor analysis.**
Note. The number of components used for clustering is indicated by j^*.

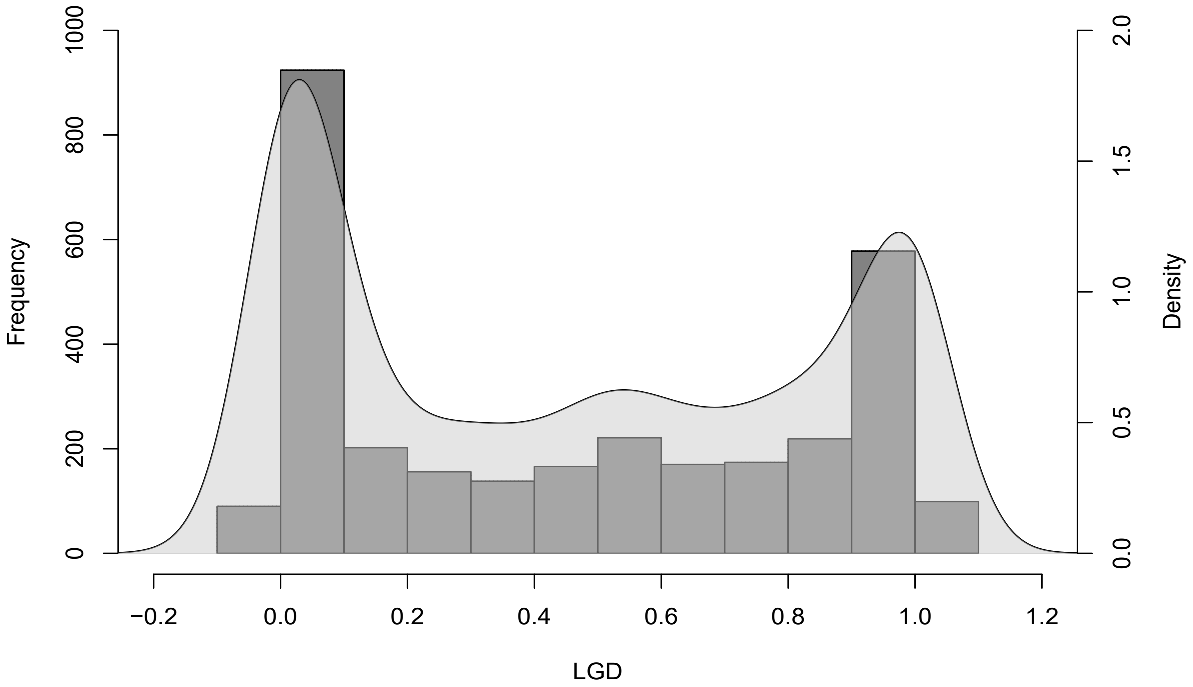

***Figure OA.2*. EU data: LGD frequency and approximated density distribution.**

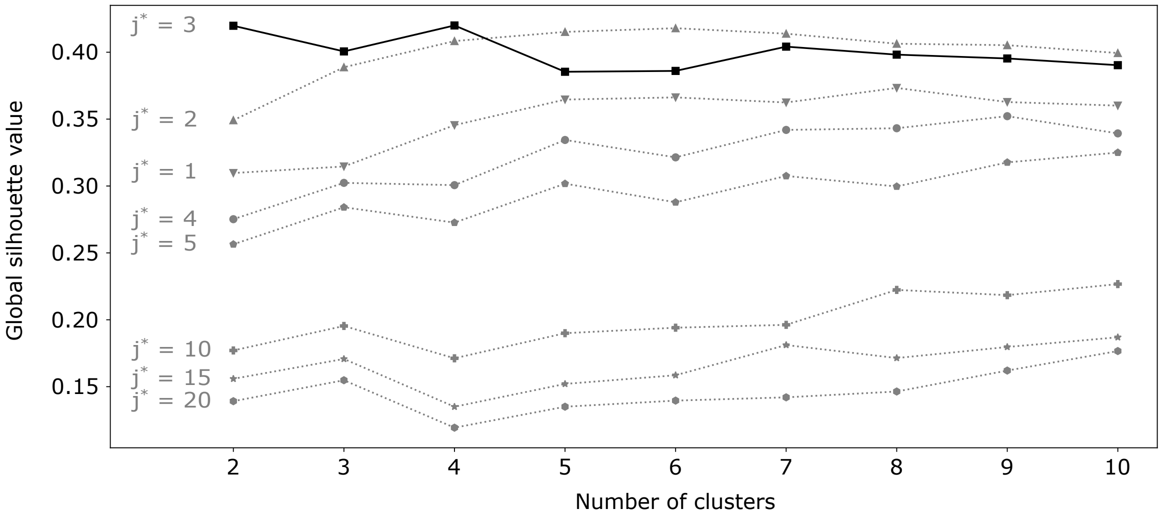

***Figure OA.3*. EU data: Clustering results of the clustering model using factor analysis.**
Note. The number of components used for clustering is indicated by j^*.

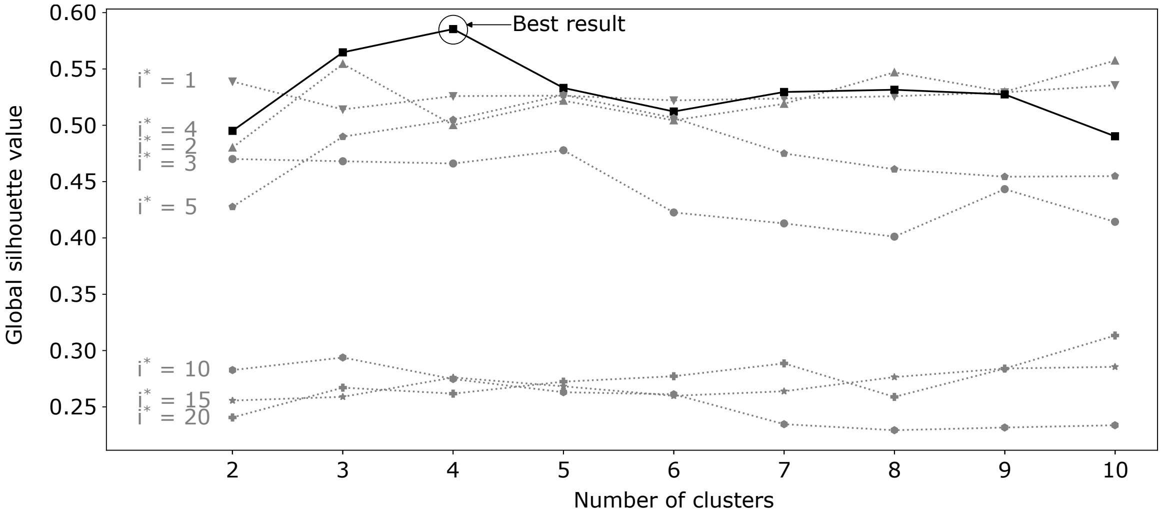

***Figure OA.4*. EU data: Clustering results of the optimized clustering model.**
Note. The number of variables used for clustering is indicated by i^*.

Tables

Table OA.1

EU data: Descriptive statistics.

Note. We consider a selected credit portfolio with defaulted loans by enterprises from Czech Republic, Denmark, Lithuania, Norway, Poland, and Romania. The table shows the means and quantiles of loan characteristics, macroeconomic factors, and of empirical LGDs (in %) for various loan categories.

Variable	Level	Quantiles					Mean	Obs.
		0.05	0.25	0.50	0.75	0.95		
$\text{LGD}_{overall}$		0.10	4.20	43.54	85.02	100.00	45.56	3137
log(EAD)		9.24	11.12	12.31	13.42	15.33	12.28	3137
Number of collaterals		0.00	0.00	1.00	1.00	4.00	1.17	3137
Number of guarantors		0.00	0.00	0.00	0.00	1.00	0.11	3137
LGD conditional to loan categories:								
Guarantee indicator	No guarantee	0.14	4.53	46.49	87.05	100.00	46.94	2866
	Guarantee	−3.01	1.40	15.18	60.10	98.29	31.02	271
Collateral indicator	No collateral	0.35	4.96	55.81	98.72	100.00	52.94	1086
	Collateral	0.03	3.77	37.02	76.44	100.00	41.66	2051
Facility type	Medium term	0.08	4.39	37.83	81.12	100.00	43.19	2273
	Short term	0.15	3.50	55.09	94.99	100.00	51.80	864
Seniority type	Pari-passu	0.23	3.44	41.90	93.58	100.00	46.64	2043
	Super senior	0.00	7.77	47.65	76.10	96.86	44.96	935
	Non senior	0.13	5.15	26.05	60.32	90.65	35.30	159
Facility asset class	Small/Medium	0.13	4.77	49.44	87.29	100.00	47.90	2782
	Large	−1.17	2.85	10.26	47.00	100.00	27.28	355
Syndication indicator	No syndication	0.12	4.24	43.37	85.28	100.00	45.64	3102
	Syndication	−7.09	−0.30	51.53	53.41	99.08	38.71	35
Lender limit	No Limit	0.24	4.13	46.74	92.68	100.79	47.57	1376
	Limit	0.04	4.25	42.77	79.23	100.00	44.00	1761
Borrower type	Public	−5.95	0.97	16.65	49.14	97.07	26.92	110
	Private	0.12	4.39	45.36	85.99	100.00	46.24	3027
Industry type								
Finance, insurance, real estate	(FIRE)	0.03	2.63	23.90	51.83	100.00	31.90	128
Agriculture, forestry, fishing, hunting	(AFFH)	0.93	6.85	60.62	97.02	100.82	54.15	495
Mining	(MIN)	0.42	12.29	35.66	58.67	99.66	39.56	25
Construction	(CON)	0.11	2.34	40.52	82.38	100.00	43.22	315
Manufacturing	(MAN)	0.08	3.44	41.65	89.66	100.00	45.48	528
Transp., commu.,elec., gas, sani. serv.	(TCEGS)	−1.45	3.84	38.55	78.27	100.00	43.35	209
Wholesale and retail trade	(WRT)	0.08	5.43	42.53	83.52	100.00	45.50	716
Services	(SERV)	0.39	12.80	53.43	89.76	100.00	52.21	297
Other	(Other)	0.01	3.22	30.82	68.78	99.99	38.40	424
STOXX 600 (rel. change)	(STOXX600)	−43.76	−23.07	5.35	16.28	30.65	−2.06	3137
STOXX 50 (rel. change)	(STOXX50)	−43.85	−25.19	−0.21	15.12	29.82	−4.69	3137
6-month EURIBOR (abs. spread in p. p.)	(6M-EURIBOR)	0.34	1.06	2.04	2.99	4.77	2.15	3137
12-month EURIBOR (abs. spread in p. p.)	(12M-EURIBOR)	0.54	1.32	2.12	3.13	4.75	2.32	3137
10-year bond yield (abs. spread in p. p.)	(EU10Y)	2.68	3.73	4.09	4.33	5.03	3.97	3137
GDP growth rate (annual %)	(GDP)	−0.07	0.00	0.03	0.04	0.06	0.01	3137
Unemployment rate (annual %)	(UNEMP.)	7.47	8.72	9.26	10.16	11.89	9.49	3137
Economic sentiment index	(ESI)	72.00	89.10	95.80	100.80	108.62	93.76	3137

Table OA.2
US data: Description of explanatory variables.

Variable	Level	Symbol	Definition
Exposure at default	-	log(EAD)	The logarithmized loss exposure (balance at the time of default) for a bank when a debtor defaults on a loan.
Number of collaterals	-	No. collaterals	Number of collaterals deposited with the loan.
Number of guarantors	-	No. guarantors	Number of guarantee commitments deposited with the loan.
Collateral indicator	No collateral Collateral	Coll.Ind.	In loan agreements, collateral is a borrower's pledge of specific property to a lender to secure repayment of a loan.
Guarantee indicator	No guarantee Guarantee	Gua.Ind.	Promise by one party (the guarantor) to assume the debt obligation of a borrower if that borrower defaults.
Facility type	Medium Short	Facility	Contractually fixed term of the loan.
Seniority type	Pari-passu Non-senior Super senior	Seniority	Seniority refers to the order of repayment in the event of default by the borrower. Senior payment obligations must be repaid before subordinate obligations.
Facility asset class	Small/Medium Large	Asset class	Type of borrower. Borrowers in the small/medium corporate asset class are defined in §218 and § 273 Basel II Accord, where the reported sales for the consolidated group of which the firm is a part is less than $50 million. For large borrowers, the reported sales for the consolidated group of which the firm is a part is above or equal to $50 million.
Syndication indicator	No syndication Syndication	Syndication	Indicates if the loan is part of a syndication, that is, if the loan is extended by a group of financial institutions (a loan syndicate) to a single borrower.
Lender limit	No limit Limit	Limit	Variable indicating whether the customer has been notified of a limit or whether a maximum amount that the bank is willing to grant the customer for a contingent facility has been approved. If the information is not available (unknown) for individual loans, these are treated as loans without a limit.
Borrower type	Public SPV Private	Borrower	Variable giving insights into the type of defaulted borrower. Public borrowers are defined as publicly listed or state-owned entities and their wholly-owned subsidiaries. Special Purpose Vehicles are specifically created for loan agreements and controlled by a few parties. In the case of private borrowers, the equity is fully in private hands.
Industry type	FIRE, AFFH MIN, CON MAN, TCEGS WRT, SERV, Other	Industry	Variable capturing the industry in which the defaulted borrower operated.

Continued on the next page

Table OA.2
(continued.)

Variable	Level	Symbol	Definition
S&P500	-	S&P500	Relative year-on-year growth rate of the S&P 500, which is the rate of change expressed over the corresponding period (month) of the previous year.
3-month LIBOR	-	LIBOR	London Interbank Offer Rate, which is the global reference rate for unsecured short-term borrowing in the interbank market. The three-month USD LIBOR is used.
Term spread	-	Term	The term spread is the difference between interest rates of short- and long-dated government securities. The difference between the 10-year and three-month government bond yields is used.
TED spread	-	TED	The TED spread is the difference between the interest rates on interbank loans and short-term US government debt ("T-bills"). The difference between a three-month LIBOR based on US dollars and three-month Treasury Bill is used.
10-year bond yield	-	US10Y	10-year government bond yield from the US.
CBOE volatility index	-	VIX	It is a real-time index that represents the market's expectations for the relative strength of near-term price changes of the S&P 500 Index.
GDP growth rate	-	GDP	The annual average rate of change in the gross domestic product at market prices, based on the US dollar.
Inflation rate	-	Inflation	Inflation rate in the US which is the price of the total basket in a given month compared with its price from the same month in the previous year.
Unemployment rate	-	UNEMP.	Percentage of the total labor force in the US that is unemployed but actively seeking employment and willing to work.
Consumer confidence index	-	CCI	It indicates the future developments of households' consumption and savings. A value above 100 signals consumers' positive confidence regarding the future economic situation; accordingly, they are less prone to save and more inclined to spend money on major purchases in the next 12 months. Values below 100 indicate a pessimistic attitude toward future developments in the economy, possibly resulting in a tendency to save more and consume less.
Producer price index	-	PPI	It measures the annual average change in the selling prices received by domestic producers for their output.
Consumer price index	-	CPI	It measures the annual average price change of all goods and services bought by households for consumption purposes.

Table OA.3

Hyperparameter choice for the machine learning models used as intelligent variable selection techniques in the optimized clustered model.

Note. The gradient-boosted trees are used in the optimized clustered model with U.S data. Random forest is used for the European data. The names of the chosen hyperparameters and a description of the machine learning methods can be found in Hastie et al. (2017). The hyperparameter sets are inspired by Breiman (2001), Qi & Zhao (2011), and Hastie et al. (2017)

Method	Hyperparameter	Description of hyperparameter	Hyperparameter set	Choice (MSE)	Choice (MAE)
Random forest	tree size	Tuning parameter that controls the tree's complexity and indicates how deep the tree is allowed to be.	{3, 4, ..., 100}	12	11
	min node size	The minimum number of observations required to split an internal node. By increasing the node size, the tree becomes more constrained because it has to consider more samples at each node.	{5, 6, ..., 50}	8	15
	min leaf size	The minimum number of observations remaining in the samples at the leaf node.	{5, 6, ..., 50}	16	22
	# splitting variables	The number of variables to consider when looking for the best split. It handles the so-called "bias-variance trade-off".	{1, 6, ..., 50}	8	9
	# trees	This parameter specifies the number of trees in the forest of the model and thus controls the model's complexity.	{100, 101..., 5000}	846	877
Gradient-boosted trees	tree size	Same description as for Regression Tree.	{3, 4, ..., 100}	28	30
	min node size	Same description as for Regression Tree.	{5, 6, ..., 50}	8	7
	min leaf size	Same description as for Regression Tree.	{5, 6, ..., 50}	13	19
	# splitting variables	Same description as for Random Forest.	{1, 6, ..., 50}	7	6
	# trees	Same description as for Random Forest.	{100, 101, ..., 5000}	176	197