Bosker, Joost; Gürtler, Marc; Zöllner, Marvin

Article — Published Version

Machine learning-based variable selection for clustered credit risk modeling

Journal of Business Economics

Suggested Citation: Bosker, Joost; Gürtler, Marc; Zöllner, Marvin (2024) : Machine learning-based variable selection for clustered credit risk modeling, Journal of Business Economics, ISSN 1861-8928, Springer Berlin Heidelberg, Berlin/Heidelberg, Vol. 95, Iss. 4, pp. 617-652, https://doi.org/10.1007/s11573-024-01213-8

This Version is available at:
https://hdl.handle.net/10419/323470

Journal of Business Economics (2025) 95:617–652
https://doi.org/10.1007/s11573-024-01213-8

ORIGINAL PAPER

Machine learning-based variable selection for clustered credit risk modeling

Joost Bosker[1] · **Marc Gürtler**[1] · **Marvin Zöllner**[1]

Accepted: 28 September 2024 / Published online: 3 December 2024

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, appropriate 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 based on theoretical and regulatory considerations. Formally, the most influential risk drivers from a best-in-class machine learning model are identified using Shapley values and employed 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 our proposed approach to clustered modeling leads to the highest prediction accuracy among various clustering models.

Keywords Credit risk · Forecasting · Clustering · Machine learning · Global credit data

JEL Classification C38 · C45 · C52 · C53 · G21

✉ Marc Gürtler
marc.guertler@tu-bs.de

Joost Bosker
j.bosker@tu-bs.de

Marvin Zöllner
marvin.zoellner@tu-bs.de

[1] Department of Finance, University of Braunschweig–Institute of Technology, Abt-Jerusalem-Straße 4, 38106 Braunschweig, Germany

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 and regulatory requirements for a meaningful differentiation of risk, a single statistical model is usually not sufficient to capture the risk characteristics of various individuals. Clustered[1] 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)].[2] 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), 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 and Murphy (2018)], which negatively affects the predictive performance in separate modeling. Therefore, the selection of appropriate variables used in cluster analysis is particularly challenging, especially for high-dimensional data. 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 a novel variable selection process on the basis of machine learning and Shapley values.[3] More formally, our approach is to calibrate a best-in-class[4] machine learning model and then use the Shapley values of Shapley (1953) as an importance measure to

[1] Grouping a set of objects so that the objects in the same group are more similar than the objects in other groups is a well-known procedure in the literature. For instance, see Cajias et al. (2020) and Gürtler and Zöllner (2023a).

[2] Refer to Sects. 3.1 and 3.2 for the theoretical background and the requirements to increase the predictive accuracy through clustered modeling.

[3] Machine learning algorithms are also increasingly being used in other areas of business economics. For instance, Schneider and Brühl (2023) investigate the predictive power of CEO characteristics on accounting fraud utilizing a machine learning approach. Götze et al. (2023) compare different machine learning approaches for predicting CAT bond premia in the secondary market.

[4] We rely on the literature to choose a "best-in-class" machine learning model. In particular, Lessmann et al. (2015) and Gürtler and Zöllner (2023a) show the superiority of gradient-boosted trees and random forest in credit risk modeling. Consequently, we consider these models the best in the class of methods that are extensively covered in the credit risk literature.

determine the risk drivers that most strongly affect the estimations. In the next step, the number of considered variables is determined and the most important variables 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. Consequently, we obtain an appropriate 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 to this procedure as "Shapley-based clustering". From the set of machine learning algorithms, we choose the best-in-class method to determine the Shapley values. For the main sample using US data, this is gradient-boosted trees by Friedman (2001).[5] For the robustness check using EU data, this is random forest by Breiman (2001).

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 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 Shapley-based clustering approach, we conduct an intensive benchmark study. Specifically, we apply the Shapley-based clustering 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 Shapley-based clustering approach considerably outperforms competing approaches. In this context, we find that clustering based on the three most important risk drivers for LGD leads to the best clustering on the US data, which significantly improves the out-of-sample performance of the predictive models in separate modeling. Moreover, the Shapley-based clustering 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 Shapley-based clustering approach does not depend on the choice of a particular dataset.

[5] Gürtler and Zöllner (2023a) show that the best estimation method depends on the LGD distribution and consequently on the geographical region.

This study contributes to the literature on clustered credit risk modeling by proposing a novel 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 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, 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[6], 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 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 appropriate variable selection for clustered modeling and has several advantages. First, using machine learning algorithms, the appropriate 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

[6] The regulatory background is explained in Sect. 3.3. In credit risk modeling, explainability is generally required in Article 179(1)(a) of the Capital Requirements Regulation (CRR) [see European Banking Authority (2013)].

approach suitable for a wide range of applications, such as the estimation of PDs, EADs or LGDs.

The remainder of this paper is organized as follows. In Sect. 2, we introduce our Shapley-based clustering approach based on a best-in-class machine learning algorithm for variable selection. In Sect. 3, we explain the proposed approach on the basis of theoretical and regulatory considerations. Section 4 presents the empirical data and settings used for the comparative analyses and describes the competitive modeling approaches. In Sect. 5, we compare the out-of-sample performance of all the modeling approaches using various evaluation criteria. Section 6 presents a robustness check. Finally, Sect. 7 concludes the paper.

2 Shapley-based clustering

In this section, we describe the Shapley-based clustering approach, which is schematically illustrated in Fig. 1. The approach is motivated on a theoretical and regulatory basis in Sect. 3 and empirically validated in Sects. 4, 5 and 6 using US and EU credit data. It 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 a best-in-class 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.[7] 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 error on the test set) are selected.[8] After calibrating the machine learning model, we calculate the Shapley value, originally introduced by Shapley (1953), for each explanatory variable. The concept of Shapley values comes from cooperative game theory and has an intuitive interpretation. The underlying problem that Shapley values solve is that a group of

[7] For a detailed description of the tree-based models and parameters to be determined, see, e.g., (Hastie et al. 2017, pp. 305–313).

[8] 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.

N players achieve a joint payoff $v(\{1, \ldots, |S|\})$ that must be distributed fairly among the players on the basis of a function $\varphi(v)$. The payout for each subset of players $S \subseteq N$ is given by $v(S)$. For a fair distribution $\varphi(v)$ some properties are specified: efficiency, symmetry, linearity and the null player property. It guarantees that the payout is distributed in full, that two players (j, k) with $v(S \cup \{j\}) = v(S \cup \{k\})$ for every $S \subseteq N$ get the same payout, the linearity of the payout function and that a player who never contributes to v gets no payout. The Shapley value is the only payout function that fulfils all the properties. For each player i the Shapley value $\varphi_i(v)$ is defined as the sum over all player subsets without player i weighted by the marginal gain in the payout when player i is involved $(v(S \cup \{i\}))$ compared to when player i is not involved $(v(S))$:

$$\varphi_i(v) = \sum_{S \subseteq N\{i\}} \frac{|S| \cdot (n - |S| - 1)!}{n!} (v(S \cup \{i\}) - v(S)). \tag{1}$$

The concept can be employed as a variable importance measure. Here, the players are the variables in the machine learning model and the payout is the relevant performance metric. In this way, a Shapley value can be viewed as the average contribution of a variable to the prediction of a machine learning model. We employ Shapley values to select the most important variables and utilize them for clustering the data. Because the determination of the Shapley value is highly computationally intensive for high-dimensional data, we use the Tree SHAP algorithm by Lundberg and Lee (2017), which approximates Shapley values with relatively low computational intensity.[9]

The third step involves selecting the optimal number of clusters. In this context, we first standardize[10] 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.[11] The distance d between the numeric variables x_i, x_j is evaluated based on the Euclidean distance $(x_i - x_j)^2$ and the similarity between the categorical variables z_k, z_l are evaluated on the ground of an indicator function:

$$\delta(z_k, z_l) = \begin{cases} 0 & \text{for } z_k = z_l, \\ 1 & \text{for } z_k \neq z_l. \end{cases} \tag{2}$$

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

[10] 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.

[11] We want to use a distance-based rather than a model-based clustering algorithm to ensure the interpretability of the resulting clusters. Several studies provide evidence that k-prototype is the best distance-based cluster method. For example, Preud'homme et al. (2021) find based on various real-life datasets that "in most of the tested scenarios, model-based methods and K-prototype typically performed best in the setting of heterogeneous data". Additionally, the authors performed a simulation study which "revealed the dominance of K-prototypes, Kamila and LCM models over all other methods". K-prototype is the only distance-based measure among the mentioned algorithms. The results obtained by Özlem and Yüksel (2021) and Jain et al. (2021) support these findings.

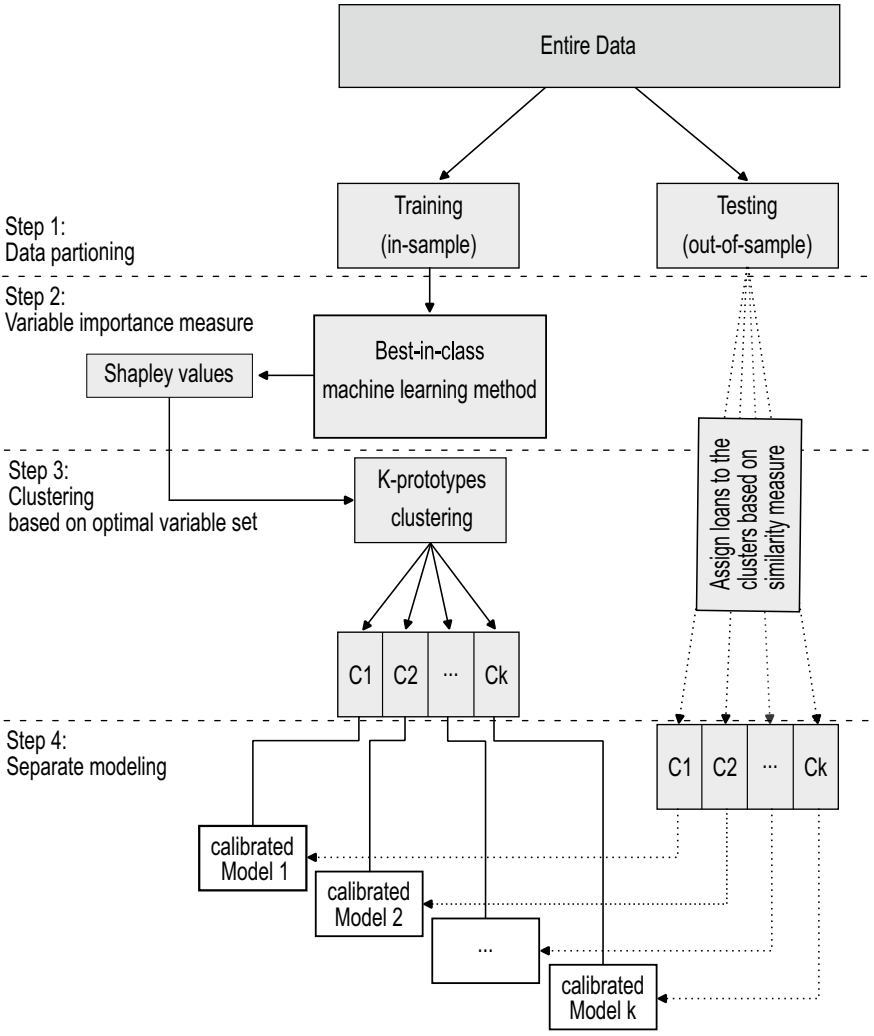

Fig. 1 Schematic diagram of the Shapley-based clustering approach

For mixed data $m^{(1)}, m^{(2)}$, where the first n variables are numeric and the last h variables are categorical, the k-prototype algorithm combines both distance measures:

$$d(m^{(1)}, m^{(2)}) = \sum_{i=1}^{n} (m_i^{(1)} - m_i^{(2)})^2 + \gamma \sum_{j=n+1}^{n+h+1} \delta(m_j^{(1)}, m_j^{(2)}). \tag{3}$$

Huang (1998) suggests to use for γ the average standard deviation of the numeric attributes.[12] 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 Shapley-based variable selection, which is the basis for clustering. The clustering process (Step 3) 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}[13] 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[14] 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 determines the final number k^* of clusters and the appropriate variable set (z_1,..., z_i^*).
- Step 3.6: Choose the k^* optimal clusters $c_1, c_2, ..., c_{k^*}$, based on the chosen 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

[12] This is to our knowledge the default configuration in all standard libraries that implement the k-prototype algorithm.

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

[14] 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.

assigned to the respective (cluster) subsamples based on the same similarity measure used in the k-prototype algorithm.[15]

For reasons of clarity and comprehensibility, we specify some important terms. In the following we consider the entire "clustered modeling approach" (Fig. 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 best-in-class machine 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 Theoretical and regulatory background

After describing our Shapley-based clustered modeling approach in Sect. 2, we motivate and explain the details of our approach. In particular, we explain why a clustering step can be advantageous and why it is important to perform the clustering only on the most important variables.

From an empirical standpoint, we observe that clustering steps are commonly employed in various research areas, such as credit risk modeling (Bakoben et al. 2020) and image recognition (Ates and Gorguluarslan 2021), to increase the predictive accuracy of machine learning models. But surprisingly, from a theoretical standpoint, it is not trivial to argue why clustering can increase the prediction accuracy of a machine learning method. On the one hand, clustering reduces the required complexity[16] of the machine learning methods to appropriately model the relationships in the data segments compared to the relationships in the non-clustered data (and consequently reduces the risk of overfitting subsamples of data). But on the other hand, the clusters contain less observations than the full sample (consequently, the risk of overfitting in the clusters increases compared to the full sample). Therefore, it is important to get a theoretical understanding of the interplay between the (required) sample size in the clusters, a method's complexity and the out-of-sample error. To this end, we first introduce some concepts from the field of statistical learning theory in Sect. 3.1 that are commonly employed to provide a theoretical basis for a machine learning approach.[17] We then apply these results to clustered risk modeling in Sect. 3.2. Finally, we discuss the influence of regulatory requirements on the empirical model design in Sect. 3.3.

[15] 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.

[16] We introduce a definition of a model's complexity, the VC dimension, in Sect. 3.1. It basically measures how deep a decision tree is or how many layers and nodes an artifical neural network contains.

[17] For example, support vector machines minimize the VC dimension among all binary classification methods with linear hyperplanes.

3.1 Statistical learning theory

First, we introduce some terms and apply them to LGD modeling. We assume that an unknown true LGD distribution $\mathcal{D}_{\text{LGD}}$ exists that (only) depends on borrower, security, bank and macroeconomic characteristics $\mathcal{X}$. Therefore, a function $f : \mathcal{X} \mapsto \mathbb{R}$ exists that maps the credit characteristics $\mathcal{X}$ to the correct LGD value. The relationship between $\mathcal{X}$ and LGD that is given by f can only be estimated on a subsample. This is the credit portfolio $\mathcal{P} = ((x_1, \text{lgd}_1), \ldots, (x_n, \text{lgd}_n)) \subset \mathcal{X} \times \mathbb{R}$ that is a sample of size n drawn randomly and independently from $\mathcal{D}_{\text{LGD}}$. A machine learning algorithm $A(\mathcal{P})$ aims to find a function h from a set of functions $\mathcal{H}$ by minimizing the empirical error (or in-sample error) $L_{\mathcal{P}}(h)$[18]:

$$L_{\mathcal{P}}(h) := \frac{\sum_{i=1}^{n} |h(x_i) - lgd_i|}{n}. \tag{4}$$

However, generally we are not interested in the empirical error, but in the true error or out-of-sample error. The true (expected) error is defined with respect to the true LGD values $f(x)$:

$$L_{\mathcal{D}_{\text{LGD}},f}(h) := E(|h(x) - f(x)|). \tag{5}$$

The true error can be decomposed into two error terms: The approximation error ϵ_{app} and the estimation error ϵ_{est}. The approximation error is the minimal error that can be achieved based on $\mathcal{H}$:

$$\epsilon_{app} = \min_{h \in \mathcal{H}} L_{\mathcal{D}_{\text{LGD}},f}(h). \tag{6}$$

The estimation error is the difference between the approximation error and the error that is achieved by minimizing the empirical error $L_{\mathcal{P}}(h)$. The error terms are influenced in opposite ways by the complexity of $\mathcal{H}$. A more complex function set $\mathcal{H}$ leads to a lower approximation error, but a higher estimation error. This is referred to as the bias-complexity trade-off.

For an artificial neural net, we can guarantee the existence of a net configuration that leads to an approximation error of zero (Ismailov 2023).[19] Consequently, various machine learning methods are able to replicate segmentation steps and achieve an approximation error of zero, but for a fixed number of observations this is only possible in-sample. Therefore, we might not be able to further reduce the approximation error of a machine learning method through clustering, but we can influence the estimation error by adjusting the complexity of the method. However, specifically for clustering we have to consider the sample size. While it is obvious that a clustering step decreases the required complexity of the machine learning method in

[18] For tree-based models, the set of functions $\mathcal{H}$ would include all tree-based functions that are representable with the chosen hyperparameter set. For instance, the machine learning algorithm A could minimize the mean absolute error in the sample.

[19] The underlying results is the general approximation theorem. But the theorem is non-constructive can not guarantee any bound on the out-of-sample performance.

each cluster, the sample size in the clusters is reduced compared to the full sample. This is a trade-off that needs to be evaluated.

To this end, two additional concepts have to be introduced: A measure for the complexity, the VC-dimension, and a measure for the required sample size, the sample complexity. We start with the sample complexity and the probably, approximately correct (PAC) learnability. The idea is that we need a minimum amount of observations in a sample to train a machine learning method with a high certainty that achieves a high (out-of-sample) accuracy. A method that achieves a high accuracy with a high probability is referred to as a PAC learner. Specifically, an algorithm A is a PAC learner if a function $n_{\mathcal{H}}(\epsilon, \delta)$ exists that leads to a true error below ϵ with a probability larger than $1 - \delta$ as long as the sample size is larger than $n_{\mathcal{H}}(\epsilon, \delta)$. The function $n_{\mathcal{H}}(\epsilon, \delta)$ is referred to as the sample complexity and is essentially the required sample size to get a true error below ϵ.

Next, we connect the sample complexity with a measure for the complexity of a model. We utilize the VC dimension as a measure for the complexity of a machine learning method (Vapnik and Chervonenkis (1971)). In its original form, it is only applicable to binary classification tasks, but the concept can be extended to real value functions with Pollard's pseudo dimension.[20] The VC dimension refers to the maximum number of clusters that an algorithm can segment the data in. As an example, assume that we have an algorithm that places a hyperplane in the $\mathbb{R}^2$ space to split the data into two classes. The algorithm has a VC dimension of two since it can cluster the data into two segments (above and below the hyperplane). It can be shown that the VC dimension is closely related to the sample complexity (Hanneke 2016; Ehrenfeucht et al. 1989)[21]:

$$\max\left\{\frac{VC(\mathcal{H})-1}{32\epsilon}, \frac{1-\epsilon}{\epsilon} ln\left(\frac{1}{\delta}\right)\right\} \leq n_{\mathcal{H}}(\epsilon,\delta) \leq c\frac{VC(\mathcal{H}) + log\left(\frac{1}{\delta}\right)}{\epsilon} \qquad (c > 0). \tag{7}$$

Therefore, a lower and an upper bound of the sample complexity depends linearly on the VC dimension. Consequently, when the complexity $VC(\mathcal{H})$ is reduced by a factor, the sample size n can be reduced by nearly the same factor without violating condition (7) as long as the sample size does not fall below the second lower bound $\frac{1-\epsilon}{\epsilon} ln\left(\frac{1}{\delta}\right)$.

[20] Pollard's pseudo dimension is beyond the scope of this article. Therefore, we focus on the more intuitive VC dimension.

[21] There are more precise characterizations of the dependence between $n_{\mathcal{H}}$ and $VC(\mathcal{H})$, which can be represented using Landau's symbols. The rough representation chosen here should suffice for the basic idea.

3.2 Application of statistical learning theory to clustered risk modeling

To apply the theoretical results to clustered risk modeling, we take the example of a binary decision tree algorithm (DT).[22] For the following considerations, we want to repeat two results from Sect. 3.1. First, according to (7), a reduction of VC leads to a smaller lower bound on the sample complexity. Therefore, the true error ϵ can be reduced without violating condition (7). Consequently, a reduction in the VC dimension generally leads to a reduced ϵ.[23] Second, we stated that when VC is reduced by a factor, the sample size n can be reduced by nearly the same factor without violating condition (7). This implies the obvious result that for fixed VC and reduced n the error ϵ tends to increase. To explain the further relationships in more detail, we use DT_0 for a decision tree trained on the non-clustered data and DT_k for a decision tree trained on cluster k. Furthermore, CA is the clustering approach, CA+DT is the combination of clustering and the use of a decision tree in each cluster and CA+DT_k is the segmentation performed by CA in combination with a specific DT_k.

We first observe that both the CA and a DT split the data: A k-means algorithm splits the data based on the euclidian distance to each cluster centrum and DT splits the data based on iteratively defining thresholds on individual variable values based on an information gain criterium. DT ends to split a node if not enough observations (defined by a hyperparameter) remain in the node or if the maximum tree size (again, defined by a hyperparameter) is reached. The number of resulting clusters from the splits corresponds to the VC dimension. Precisely, the VC dimension of DT is the number of leaf notes and the VC dimension of CA is the number of clusters.

We introduce two simplifications to present the following in a concise manner. We assume that the dependent variable LGD is always either zero or one and that both CA+DT and DT_0 can and will split the training data so that only LGD = 0 or LGD = 1 remain in each cluster. Essentially, we are assuming that CA+DT and DT_0 both have an approximation error of zero for a fair comparison.

To assess whether CA+DT can achieve a lower VC dimension than DT_0, the resulting numbers of clusters must be compared that are required to fully separate the data based on LGD. Precisely, the number of the resulting clusters from the splits performed by DT_0 have to be compared with the combined number of the resulting clusters from the splits performed by each CA+DT_k. Since closed formulas are generally not derivable, we will look at two extreme cases and give an insight into the reasons for the different results.

The first example is depicted in Fig. 2. The two variables X_1 and X_2 can perfectly be separated into the cases $Y = 0$ and $Y = 1$ by a diagonal hyperplane. A k-means algorithm can easily achieve this separation based on two clusters. DT_0 requires more than four clusters to separate the data into $Y = 0$ and $Y = 1$. Additionally,

[22] The use of a binary tree is only to simplify the notation. The same logic can also be applied to other machine learning methods. Essentially, only the number of leaf nodes are relevant and this number depends on the number of splits per node.

[23] Of course, we have to assume a constant approximation error. But this is the case since we never change $\mathcal{H}$.

depending on the chosen hyperparameters (in particular the minimum number of observations in each leaf node), DT_0 may not be able to fully separate the data.

Therefore, the VC dimension of CA+DT is two and the VC dimension of DT_0 is more than four. Since the higher number of clusters in DT_0 results in smaller clusters with less data, this method in turn is more exposed to possible overfitting than CA+DT. We conclude from the example that there are cases where the VC dimension can be reduced by employing CA+DT compared to DT_0. Here, CA+DT is preferred since CA can better model the diagonal hyperplane than DT_0.

In the second example, we randomly assign contracts to k clusters with at most $\lceil n/k \rceil$ observations. In each cluster, the observations are drawn from the same distribution $\mathcal{D}_{\text{LGD}}$. Each DT_k generally has to employ the same splits as DT_0 to achieve the same error (without proof, this is at least true for $n \to \infty$). Therefore, the VC dimension and sample complexity for each DT_k are approximately equal to the VC dimension and sample complexity of DT_0, but the number of observations in each cluster is reduced by approximately a factor of $1/k$. Consequently, the risk of overfitting increases in the clusters and ϵ is increased.

Against this background, we can conclude that CA+DT can be beneficial compared to DT_0, but it strongly depends on the clustering approach and the structure of the underlying data. To summarize, we can highlight some general aspects: First, clustering should only be performed for the most important variables. It is not beneficial to cluster data that are not correlated with LGD. This can be achieved with our Shapley-based clustering approach. Second, CA+DT tends to be better than DT_0 when the data are "diagonally" split and there are distinct clusters. Third, CA+DT may be favorable compared to DT_0, but this needs to be confirmed empirically for the specific data.

3.3 Regulatory background

Approaches in credit risk modeling can not solely be based on theoretical consideration, but have to incorporate regulatory requirements. Therefore, some peculiarities of the regulatory background and the approval process by central banks have to be discussed. Central banks base the approval process of a banks credit risk calculation and reporting on the Basel Framework and guidelines by local banking authorities. The Basel Framework specifies credit risk calculation, disclosure requirements and the supervisory review process. Generally, three approaches are available for a bank's risk reporting (depending on the asset class): the standardised approach (SA), the foundation internal ratings-based approach (F-IRB) and the advanced internal ratings-based approach (A-IRB).[24] With the F-IRB and A-IRB approaches, the banks must (partially) calculate the LGDs themselves. While the F-IRB approach provides different supervisory specified LGDs for various clusters of data,[25] the A-IRB approach only provides LGD parameter floors and fully relies on the banks

[24] See p.5 in bis.org/bcbs/publ/d424_hlsummary.pdf.

[25] See CRE32.6, CRE32.7 and CRE32.11 in https://www.bis.org/basel_framework/chapter/CRE/32.htm.

internal models.[26] The requirements to adopt an IRB approach are defined in CRE36 of the Basel Framework. Generally, LGDs have to be reported separately for different asset classes.[27] Banks have to ensure that calculated LGDs are not lower than the long-term default-weighted average loss rates for all asset classes.[28] To enable banks to appropriately estimate risk measures for each asset class, banks have to assign contracts to homogenous pools and demonstrate that the approach "provides for a meaningful differentiation of risk... and allows for accurate and consistent estimation of loss characteristics at pool level".[29] On pool level, banks are allowed to employ different rating systems and methodologies.[30] They must state and document the rationale for allocating the contracts to the individual rating systems.

Additionally, the European Banking Authority states in Article 179 of its Capital Requirements Regulation that estimates shall be plausible and intuitive (i.e, interpretable) and that the estimates have to reflect technical advances.[31] Therefore, clustered credit modeling has to be based on intuitive risk drivers and interpretable clusters. Consequently, clustering has to be performed on a subset of interpretable risk drivers. The present Shapley-based clustering approach relies (in contrast to, e.g., clustering on all available information) on only the most important risk drivers and produces distinct and interpretable clusters. Additionally, no manual and subjective process is required to cluster the data.

From a regulatory perspective, the approach appears to be in line with regulatory requirements and we are convinced that it can be approved in its entirety as consistent clustering plays an important role in the regulatory approval process. From a practical perspective, banks have to evaluate whether the increase in overhead justifies better predictions. The overhead to maintain different clusters of data is for most banks likely in place to ensure regulatory compliance. Nevertheless, the implementation of our approach requires a significant effort to setup the system. In addition, it is necessary to monitor the cluster results over time and explain when the model's determined clusters change. But an enhanced approach to cluster the data is beneficial for banks. The European Banking Authority requires that the less information a bank can utilize, the more conservative the estimations have to be,[32] which in turn leads to higher capital requirements. Additionally, the effect of biased LGDs are non-symmetric since rating floors are provided. Therefore, biased LGDs are likely to a bank's disadvantage. Finally, accurate estimations are important for internal risk models and provide a clear competitive advantage.

[26] See CRE32.16 in https://www.bis.org/basel_framework/chapter/CRE/32.htm.

[27] E.g., in the yearly stress testing by the European Central Bank. The template can be retrieved from here: lnk.tu-bs.de/B4D9yc.

[28] See CRE36.83 in bis.org/basel_framework/chapter/CRE/36.htm.

[29] See CRE36.16 in bis.org/basel_framework/chapter/CRE/36.htm.

[30] See CRE36.10 in bis.org/basel_framework/chapter/CRE/36.htm.

[31] See European Banking Authority (2013).

[32] See European Banking Authority (2013).

4 Empirical framework

To demonstrate the effectiveness of our Shapley-based clustering approach, we conduct an intensive benchmark study. For this purpose, we use the Global Credit Data[33] 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.

4.1 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,

$$LGD = 1 - \frac{\sum F^+ - \sum C^-}{EAD}. \tag{8}$$

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 and 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

[33] See https://www.globalcreditdata.org.

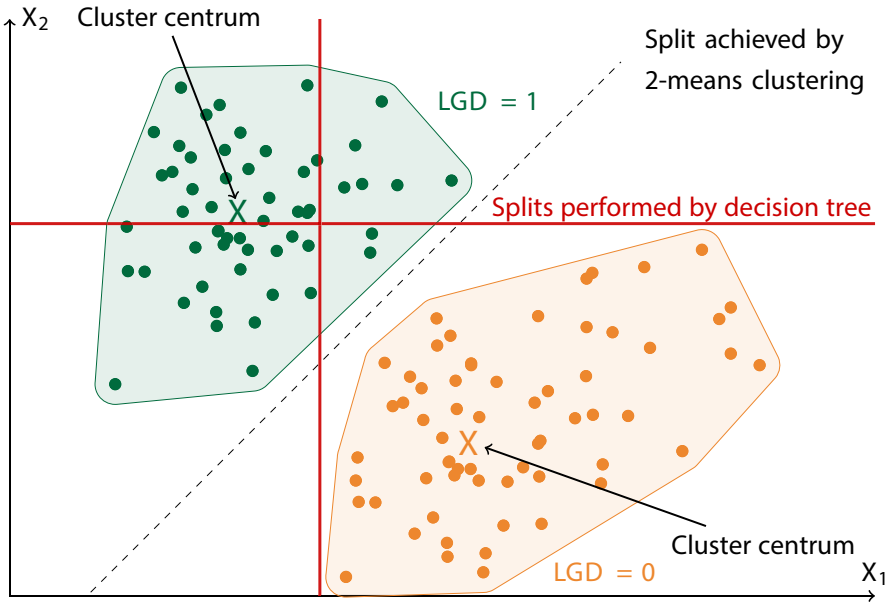

Fig. 2 CA+DT requires two clusters (round framed areas) to completely separate the data based on *Y*. DT_0 is not able to completely separate the data with four clusters (separated by the drawn horizontal and vertical line)

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 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 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.[34] Stock exchange performances are identified as general LGD risk drivers, for instance, by Qi and 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

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

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.[35]

Figure 3 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 explain why gradient-boosted trees are the best-in-class estimation method that is consequently appropriate for our selection model for clustering.[36]

4.2 Competitive modeling approaches

Based on the regulatory requirements stated in Sect. 3.3, a bank likely clusters data based on natural segmentations in the data (e.g., different asset classes).[37] Therefore, for a fair comparison, we must evaluate the effectiveness of our Shapley-based clustering approach with various clustered approaches using different clustering models.[38] 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 Fig. 1 (see Sect. 2) but differ in the selection of variables for clustering. In

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

[36] The link between the distribution type and the best estimation method has been shown by Gürtler and Zöllner (2023a). Gradient-boosted trees best capture the bimodality and the specific skewness.

[37] We observe in Sect. 5.2 that the competitive modeling approaches consider different natural clusters and are therefore a realistic comparison. For the example of SME's and LC's, most models either cluster based on the exposure or directly on the facility asset class.

[38] 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. We already argued in Sect. 2 that the k-prototype algorithm is appropriate for mixed data and the preferred approach compared to other distance-based clustering methods (Preud'homme et al. 2021).

total, we consider five different clustering models based on recommendations from the literature[39] 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 and Massicotte (2021)]. Third, we use a clustering model that employs linear regression with the *k*-best algorithm and F scores to select the variables for clustering. We calculate the correlations between each variable and the LGD and convert the correlations to F scores. We add the variables with the *k* highest F scores to a linear regression model. Then, we perform a grid search with a 5-fold cross-validation to determine the optimal number *k* of variables in the linear regression model that leads to the best cross-validation performance. These variables are then 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.[40] 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 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.

[39] 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.

[40] It is important that the variables show both statistical and economic significance to conclude that they are important. The effects of statistically significant variables can be generalized to the whole statistical population. And economic significance ensures that the correlations are large given that the variables are statistically significant. In our case, our chosen variables show both significances. The variables have clear economic interpretations and are highly correlated with the LGD.

The best hyperparameter values for the machine learning models are determined in the same manner as described in Sect. 2.[41]

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 by Loterman et al. (2012) and Krüger and 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 and 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.[42]

4.3 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 testing. For the training dataset, we use data

[41] 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.

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

Table 1 Descriptive statistics

Variable	Level	Quantiles					Mean	Obs.
		0.05	0.25	0.50	0.75	0.95		
$LGD_{overall}$		−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

Table 1 (continued)

Variable	Level	Quantiles					Mean	Obs.
		0.05	0.25	0.50	0.75	0.95		
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

This table shows the means and quantiles of the loan characteristics, macroeconomic factors, and empirical LGDs (in %) for various loan categories

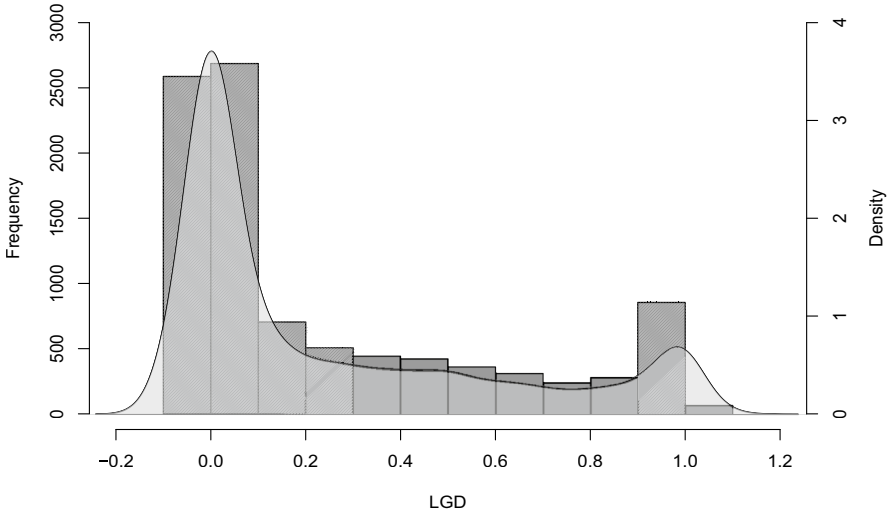

Fig. 3 LGD frequency and approximated density distribution

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:

$$MSE := \frac{1}{n}\sum_{i=1}^{n}(LGD_i - \widehat{LGD}_{i,m})^2, \tag{9}$$

$$MAE := \frac{1}{n}\sum_{i=1}^{n}|LGD_i - \widehat{LGD}_{i,m}|, \tag{10}$$

$$MedAE := median(|LGD_1 - \widehat{LGD}_{1,m}|, ..., |LGD_n - \widehat{LGD}_{n,m}|), \tag{11}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(LGD_i - \widehat{LGD}_{i,m})^2}{\sum_{i=1}^{n}(LGD_i - \overline{LGD})^2}, \tag{12}$$

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 (precisely, the mean squared error) 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).[43] 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.

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

5 Empirical results

As already mentioned, we use gradient-boosted trees as the variable selection method within the Shapley-based 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 Shapley-based clustering model is described in more detail. Finally, we state and evaluate the results of the comparative out-of-sample analyses.

5.1 Variable importance measure

After calibrating the gradient-boosted trees within the Shapley-based clustering model, we identify the most influential variables for estimating the LGD. Figure 4 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 and de Carvalho (2006), Grunert and Weber (2009), Krüger and Rösch (2017) and Betz et al. (2018)] and can be summarized as follows:

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.

5.2 Cluster analysis

Figure 5 shows the clustering results of the Shapley-based 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 6 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 the best 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 Shapley-based clustering model. In summary, all other clustering models using the baseline methods for variable selection achieve a lower cluster quality than the Shapley-based clustering model.

To characterize the resulting clusters, we list the employed variables in the clustering models below Fig. 6. We observe that most clusters are based on only a few variables with a distinct economic interpretation. To go more depth into the resulting clusters based on the Shapley-based 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. These clusters can all be interpreted economically and most contracts can clearly be assigned to one cluster. This can be deduced from the high global silhouette score. But small groups of contracts that do not fit any cluster well still have to be assigned to an ill-fitting cluster since a machine learning method can only reliably be learned on enough observations. This problem of forced cluster assignments and the possibility of a bad LGD estimate for these small groups of contracts is partly mitigated from a regulatory perspective: The reported LGDs cannot be under a predetermined LGD floor and the bank has to show that long-running historical realized LGDs are below the reported LGDs (see Sect. 3.3). But still, a bank should consider conservative LGD estimates for ill-fitting contracts to ensure regulatory compliance. But this is not limited to our clustering approach, but

to every clustered and non-clustered model: If the number of observations is too low to reliably estimate the LGD, some form of conservatism is likely required.

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 (Fig. 3). Therefore, we can confirm that the Shapley-based clustering model leads to economically meaningful and comprehensible clusters, as required by regulators.

5.3 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.

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 Shapley-based clustering approach.

Fifth, the most important result of the comparative analysis is that the Shapley-based clustering 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 Shapley-based clustering approach are also robust against outliers in the prediction errors. Additionally, even the performance of the simple linear regression is remarkably improved in the Shapley-based clustering 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 Shapley-based 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 based on machine learning and Shapley values. Additionally, we find that clustering on a sub-sample of the variables, leads to the best out-of-sample predictive performance.

6 Robustness check

To ensure that the superiority of the Shapley-based clustering 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 3137 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 and 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 a variable selection method in the Shapley-based clustering approach.

Figure 7 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.

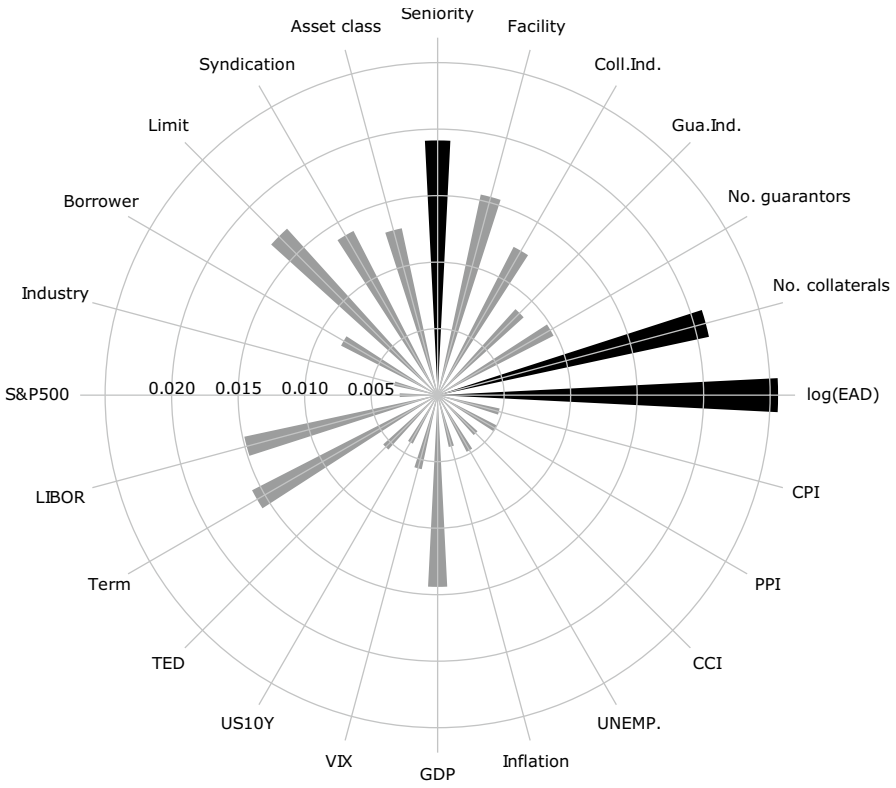

Fig. 4 Variable importance measure in gradient-boosted trees used within the Shapley-based clustering model

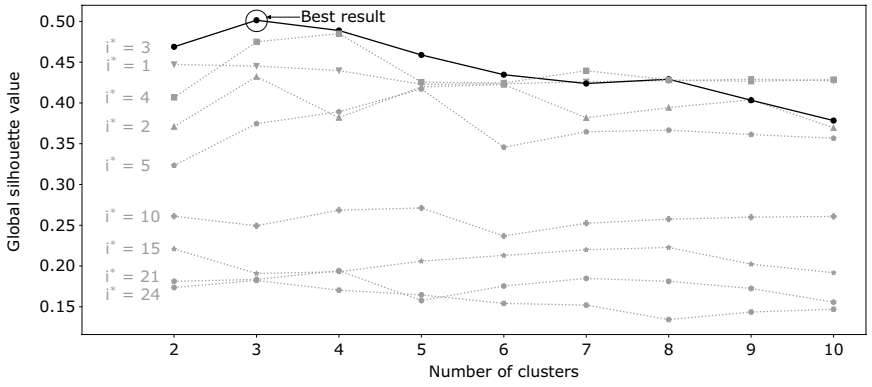

Fig. 5 Clustering results of the Shapley-based clustering model. The number of important variables used for clustering is indicated by i^*

The comparison of the clustering results of the competitive clustering models is

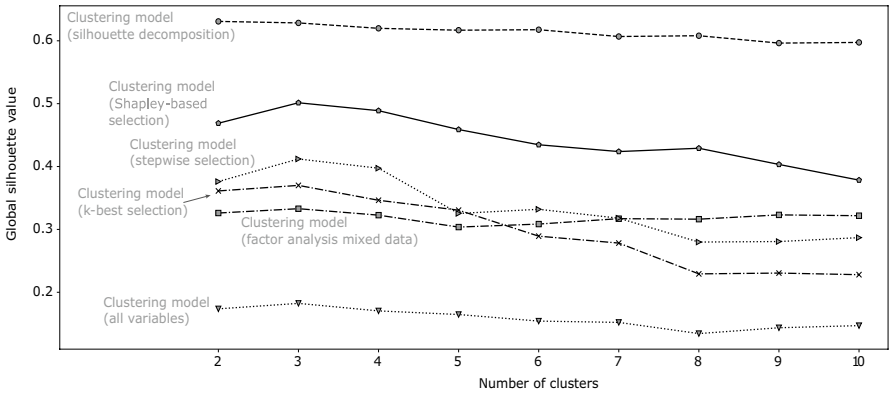

Fig. 6 Comparison of the clustering results of the competing clustering models. The variable selection methods used in the clustering models are indicated in parentheses. Additionally, the variables used in clustering are specified for each model. Clustering model (all variables): All variables Clustering 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 (Shapley-based selection): log(EAD), No. collaterals, Seniority

shown in Fig. 8.[44] 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 Shapley-based clustering 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 Shapley-based clustering approach.

Table 4 presents the mean values of the numerical variables and modal values of the categorical variables for each cluster created using the Shapley-based clustering 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 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

[44] The clustering results of the Shapley-based clustering approach and clustered approach with factor analysis are shown in detail in Figures QA.3 and QA.4 in the online appendix.

Table 2 Interpretation of the resulting clusters of the Shapley-based clustering model

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

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

model within the Shapley-based clustering 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 Shapley-based clustering approach and the need for clustering on a sub-sample of variables—can also be drawn for European data.

7 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

Table 3 Results of the comparative out-of-sample analysis

Variable selection (clustering)	LR	RT	RF	GBT	SVR	MLP
Panel A: Mean squared error (MSE)						
No clustering	0.1245	0.1230	0.1122*	0.1099*	0.1162	0.1193
All variables	0.1202	0.1195	0.1077	0.1042*	0.1113	0.1092
Silhouette decomposition	0.1197	0.1186	0.1058	0.1038*	0.1092	0.1094
K-best selection	0.1132	0.1121	0.0984	0.0971*	0.1019	0.0998
Stepwise selection	0.1102	0.1094	0.0962*	0.0949*	0.0979	0.0974
Factor analysis	0.1114	0.1102	0.0961*	0.0952*	0.0994	0.0972
Shapley-based selection	0.1085	0.1004	0.0921	0.0887*‡	0.0933	0.0914
Panel B: Mean absolute error (MAE)						
Variable selection (clustering)	LR	RT	RF	GBT	SVR	MLP
No clustering	0.3032	0.2899	0.2750*	0.2745*	0.2784	0.2863
All variables	0.2957	0.2823	0.2711	0.2686*	0.2735	0.2728
Silhouette decomposition	0.2876	0.2782	0.2678	0.2622*	0.2692	0.2732
K-best selection	0.2759	0.2655	0.2648	0.2614*	0.2655	0.2649
Stepwise selection	0.2727	0.2625	0.2615*	0.2582*	0.2613	0.2631
Factor analysis	0.2734	0.2631	0.2611*	0.2589*	0.2623	0.2625
Shapley-based selection	0.2661	0.2615	0.2553	0.2501*‡	0.2587	0.2538
Panel C: $\lvert MAE - MedAE \rvert$						
Variable selection (clustering)	LR	RT	RF	GBT	SVR	MLP
No clustering	0.0650	0.0313*	0.0341	0.0312*	0.0367	0.0567
All variables	0.0633	0.0212*	0.0290	0.0244	0.0312	0.0255
Silhouette decomposition	0.0538	0.0159	0.0246	0.0125*	0.0255	0.0262
K-best selection	0.0387	0.0154	0.0124*	0.0115*	0.0186	0.0168
Stepwise selection	0.0328	0.0109	0.0093*	0.0089*	0.0131	0.0143
Factor analysis	0.0345	0.0120	0.0080*	0.0088*	0.0145	0.0132
Shapley-based selection	0.0219	0.0150	0.0020*	0.0017*‡	0.0074	0.0031
Panel D: Coefficient of determination R^2						
Variable selection (clustering)	LR	RT	RF	GBT	SVR	MLP
No clustering	0.1217	0.1499	0.1867	0.1883*	0.1782	0.1723
All variables	0.1423	0.1727	0.2123	0.2292*	0.1989	0.2134
Silhouette decomposition	0.1689	0.1791	0.2286	0.2311*	0.2121	0.2110
K-best selection	0.1816	0.1843	0.2424*	0.2463*	0.2388	0.2373
Stepwise selection	0.2048	0.2072	0.2492	0.2555*	0.2477	0.2482
Factor analysis	0.1982	0.2036	0.2487	0.2544*	0.2381	0.2458
Shapley-based selection	0.2187	0.2389	0.2736	0.2798*‡	0.2715	0.2744

Each modeling approach is a combination of a clustering model and prediction model. The prediction models are linear regression (LR), regression trees (RT), random forest (RF), gradient-boosted trees (GBT), support vector regression (SVR) and multilayer perceptron (MLP). The employed variable selection method for the clustering is given in the first column. 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

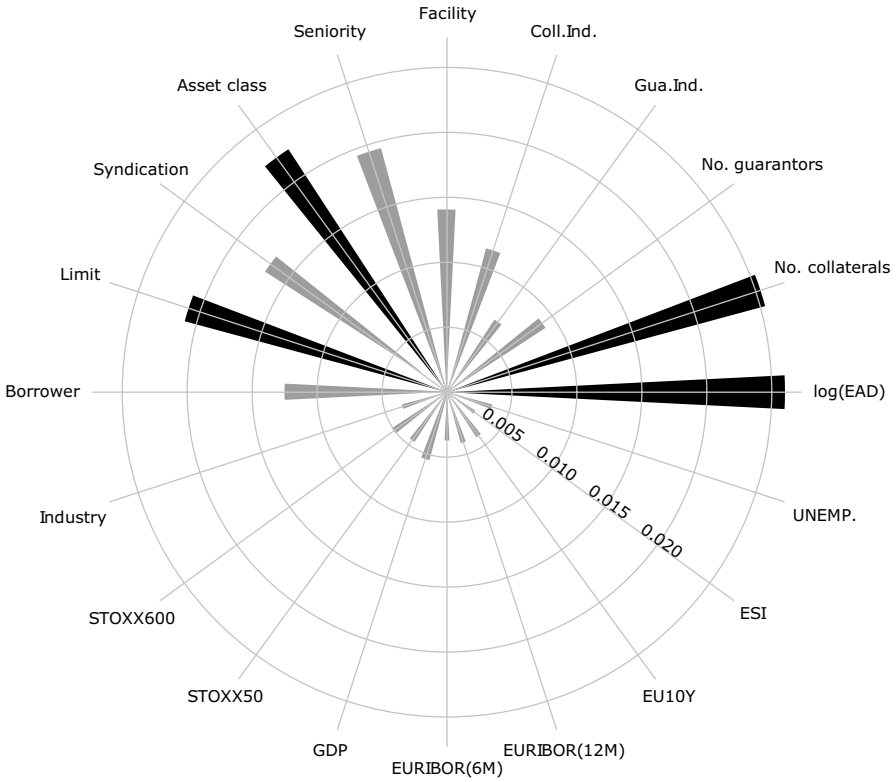

Fig. 7 EU data: variable importance measure in random forest used within the Shapley-based clustering model

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 appropriate 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 a variable selection process 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.

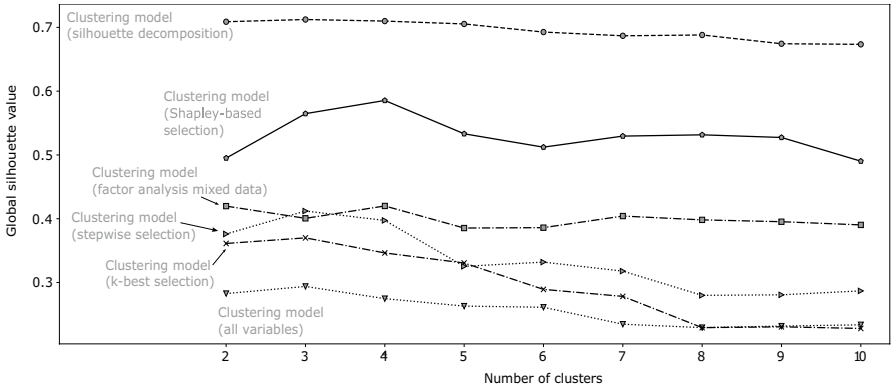

Fig. 8 EU data: comparison of the clustering results of the competing clustering models. The variable selection method used in the clustering model is indicated in parentheses. Additionally, the variables used in clustering are specified for each model. Clustering model (all variables): All variables 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(6 M), UNEMP. Clustering model (Shapley-based selection): Log(EAD), No. collaterals, Asset class, Limit

Table 4 EU data: interpretation of the resulting clusters of the Shapley-based clustering model

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(6 M)	2.01	2.20	2.50
EURIBOR(12 M)	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

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

Table 5 EU data: results of the comparative out-of-sample analysis

Variable selection (clustering)	LR	RT	RF	GBT	SVR	MLP
Panel A: Mean squared error (MSE)						
No clustering	0.1239	0.1211	0.1089*	0.1152	0.1151	0.1197
All variables	0.1212	0.1189	0.1063*	0.1112	0.1112	0.1134
Silhouette decomposition	0.1207	0.1186	0.1062*	0.1102	0.1097	0.1112
K-best selection	0.1185	0.1172	0.1048*	0.1051*	0.1061	0.1055*
Stepwise selection	0.1182	0.1155	0.1029*	0.1049	0.1032*	0.1089
Factor analysis	0.1153	0.1121	0.0982*	0.1021	0.0999	0.1064
Shapley-based selection	0.1093	0.1065	0.0892*‡	0.0931	0.0922	0.0987
Panel B: Mean absolute error (MAE)						
Variable selection (clustering)	LR	RT	RF	GBT	SVR	MLP
No clustering	0.3112	0.2954	0.2858*	0.2913	0.2908	0.2931
All variables	0.2946	0.2873	0.2771*	0.2807	0.2795	0.2841
Silhouette decomposition	0.2942	0.2870	0.2766*	0.2801	0.2796	0.2832
K-best selection	0.2931	0.2844	0.2710*	0.2753	0.2767	0.2749
Stepwise selection	0.2915	0.2826	0.2687*	0.2789	0.2727*	0.2812
Factor analysis	0.2883	0.2797	0.2653*	0.2744	0.2697	0.2789
Shapley-based selection	0.2793	0.2783	0.2592*‡	0.2633	0.2627	0.2692
Panel C: $\lvert MAE - MedAE \rvert$						
Variable selection (clustering)	LR	RT	RF	GBT	SVR	MLP
No clustering	0.0679	0.0442	0.0402*	0.0316*	0.0411	0.0423
All variables	0.0465	0.0332	0.0180*	0.0209	0.0217	0.0320
Silhouette decomposition	0.0463	0.0321	0.0165*	0.0194	0.0214	0.0302
K-best selection	0.0402	0.0278	0.0116*	0.0165	0.0189	0.0167
Stepwise selection	0.0444	0.0288	0.0105*	0.0145	0.0085*	0.0184
Factor analysis	0.0372	0.0199	0.0041*	0.0131	0.0088	0.0191
Shapley-based selection	0.0184	0.0165	0.0011*‡	0.0026	0.0019	0.0058
Panel D: Coefficient of determination R^2						
Variable selection (clustering)	LR	RT	RF	GBT	SVR	MLP
No clustering	0.1221	0.1512	0.1926*	0.1844	0.1828	0.1711
All variables	0.1506	0.1765	0.2244*	0.1992	0.1995	0.1892
Silhouette decomposition	0.1473	0.1785	0.2249*	0.2049	0.2110	0.2098
K-best selection	0.1762	0.1793	0.2282*	0.2272*	0.2388	0.2268
Stepwise selection	0.1771	0.1805	0.2303*	0.2280	0.2294*	0.1928
Factor analysis	0.1811	0.1873	0.2421*	0.2381	0.2404	0.2362
Shapley-based selection	0.2112	0.2374	0.2755*‡	0.2694	0.2711	0.2564

Each modeling approach is a combination of a clustering model and prediction model. The prediction models are linear regression (LR), regression trees (RT), random forest (RF), gradient-boosted trees (GBT), support vector regression (SVR) and multilayer perceptron (MLP). The employed variable selection method for the clustering is given in the first column. 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

The superiority of our Shapley-based clustering approach is investigated through an empirical analysis using two real-life LGD datasets. We demonstrate that the Shapley-based clustering 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 Shapley-based clustering 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.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11573-024-01213-8.

Acknowledgements 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.

Funding Open Access funding enabled and organized by Projekt DEAL. No funds, grants, or other financial support was received.

Availability of data and material Non-disclosure agreement with Global Credit Data (GCD). GCD is a non-profit association. Seehttps://www.globalcreditdata.org/ for further information.

Code availability The programming code in Python of the applied models is available.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Ates GC, Gorguluarslan RM (2021) Two-stage convolutional encoder-decoder network to improve the performance and reliability of deep learning models for topology optimization. Struct Multidiscip Optim 63:1927–1950. https://doi.org/10.1007/s00158-020-02788-w

Bakoben M, Bellotti T, Adams N (2020) Identification of credit risk based on cluster analysis of account behaviours. J Oper Res Soc 71(5):775–783. https://doi.org/10.1080/01605682.2019.1582586

Bastos JA (2014) Ensemble predictions of recovery rates. J Financ Serv Res 46(2):177–193. https://doi.org/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. Eur J Oper Res 271(3):1113–1144. https://doi.org/10.1016/j.ejor.2018.05.059

Bishop CM (1995) Neural networks for pattern recognition. Oxford University Press, Oxford

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

Breiman L (2001) Random forest. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324

Cajias M, Freudenreich P, Freudenreich A, Schäfers W (2020) Liquidity and prices: a cluster analysis of the German residential real estate market. J Bus Econ 90(7):1021–1056. https://doi.org/10.1007/s11573-020-00990-2

Caruso G, Gattone SA, Fortuna F, Di Battista T (2021) Cluster analysis for mixed data: an application to credit risk evaluation. Socio-Econ Plann Sci 73:100850. https://doi.org/10.1016/j.seps.2020.100850

Chava S, Stefanescu C, Turnbull S (2011) Modeling the loss distribution. Manag Sci 57(7):1267–1287

Dermine J, de Carvalho CN (2006) Bank loan losses-given-default: a case study. J Bank Finance 30(4):1219–1243. https://doi.org/10.1016/j.jbankfin.2005.05.005

De Soete G, DeSarbo WS, Carroll JD (1985) Optimal variable weighting for hierarchical clustering: an alternating least-squares algorithm. J Classif 2(1):173–192. https://doi.org/10.1007/BF01908074

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

Ehrenfeucht A, Haussler D, Valiant L (1989) A general lower bound on the number of examples needed for learning. Inf Comput 82(3):247–261

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. 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 TB (2018) Variable selection methods for model-based clustering. Stat Surv. https://doi.org/10.1214/18-SS119

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29(5):1189–1232. https://doi.org/10.1214/aos/1013203451

Götze T, Gürtler M, Witowski E (2023) Forecasting accuracy of machine learning and linear regression: evidence from the secondary cat bond market. J Bus Econ 93(9):1629–1660. https://doi.org/10.1007/s11573-023-01138-8

Green PE, Kim J, Carmone FJ (1990) A preliminary study of optimal variable weighting in k-means clustering. J Classif 7(2):271–285. https://doi.org/10.1007/BF01908720

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

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

Gürtler M, Zöllner M (2023) Tuning white box model with black box models: transparency in credit risk modeling. SSRN Electron J. https://doi.org/10.2139/ssrn.4433967

Hanneke S (2016) The optimal sample complexity of pac learning. J Mach Learn Res 17:1–15

Hansen PR, Lunde A, Nason JM (2011) The model confidence set. Econometrica 79(2):453–497. https://doi.org/10.3982/ECTA5771

Harris T (2015) Credit scoring using the clustered support vector machine. Expert Systems with Applications 42(2):741–750. https://doi.org/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. https://doi.org/10.1016/j.jbankfin.2013.12.006

Hastie T, Tibshirani R, Friedman JH (2017) The elements of statistical learning: Data mining, inference, and prediction (Second edition, corrected at 12th printing, 2017th edn. Springer, New York, NY

Huang Z (1998) Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery 2(3):283–304. https://doi.org/10.1023/A:1009769707641

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

Ismailov V (2023) A three layer neural network can represent any multivariate function. Journal of Mathematical Analysis and Applications, 523

Jain S, Shastri A, Ahuja K, Busnel Y, Singh N (2021) Cube sampled k-prototype clustering for featured data. IEEE 18th India Council International Conference

Krüger S, Rösch D (2017) Downturn LGD modeling using quantile regression. Journal of Banking & Finance 79:42–56. https://doi.org/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. https://doi.org/10.1016/j.eswa.2020.113940
Lessmann S, Baesens B, Seow H-V, Thomas LC (2015) Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research 247(1):124–136. https://doi.org/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. https://doi.org/10.1016/j.ijforecast.2011.01.006
Lundberg SM, 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/ 8a20a8621978632d-76c43dfd28b67767-Paper.pdf
Matuszyk A, Mues C, Thomas LC (2010) Modelling LGD for unsecured personal loans: Decision tree approach. Journal of the Operational Research Society 61(3):393–398. https://doi.org/10.1057/jors.2009.67
Milligan GW (1989) A validation study of a variable weighting algorithm for cluster analysis. Journal of Classification 6(1):53–71. https://doi.org/10.1007/BF01908588
Nazemi A, Fatemi Pour F, Heidenreich K, Fabozzi FJ (2017) Fuzzy decision fusion approach for loss-given-default modeling. European Journal of Operational Research 262(2):780–791. https://doi.org/10.1016/j.ejor.2017.04.008
Preud'homme G, Duarte K, Dalleau K, Lacomblez C, Bresso E, Smaïl-Tabbone M, Girerd N (2021) Head-to-head comparison of clustering methods for heterogeneous data: a simulation-driven benchmark. Scientific Reports. https://doi.org/10.1038/s41598-021-83340-8
Qi M, Zhao X (2011) Comparison of modeling methods for loss given default. Journal of Banking & Finance 35(11):2842–2855. https://doi.org/10.1016/j.jbankfin.2011.03.011
Rousseeuw PJ (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20:53–65. https://doi.org/10.1016/0377-0427(87)90125-7
Schneider M, Brühl R (2023) Disentangling the black box around ceo and financial information-based accounting fraud detection: Machine learning-based evidence from publicly listed u.s. firms. Journal of Business Economics 93(9):1591–1628. https://doi.org/10.1007/s11573-023-01136-w
Shapley LS (1953) A value for n-person games. In: Kuhn H, Tucker AW (eds) Contributions to the theory of games, ii. Princeton University Press
Smith G (2018) Step away from stepwise. Journal of Big Data, 5(1). https://doi.org/10.1186/s40537-018-0143-6
Tanoue, Y., & Yamashita, S. (2019). Loss given default estimation: A two-stage model with classification treebased boosting and support vector logistic regression. Journal of Risk, 21(4), 19–37.https://doi.org/10.21314/JOR.2019.405
Vapnik V (1995) The nature of statistical learning theory. New York, NY: Springer. https://doi.org/10.1007/978-1-4757-2440-0
Vapnik V, Chervonenkis A (1971) On the uniform convergence of relative frequencies of events to their probabilities. Theory of Probability and its applications 2:264–280
Yao X, Crook J, Andreeva G (2015) Support vector regression for loss given default modelling. European Journal of Operational Research 240(2):528–538. https://doi.org/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. https://doi.org/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. https://doi.org/10.1016/j.ribaf.2021.101536
Özlem Akay, Yüksel G (2021) Hierarchical clustering of mixed variable panel data based on new distance. Communications in Statistics - Simulation and Computation, 50(6)