*Article*

# Modeling Recovery Rates of Small- and Medium-Sized Entities in the US

**Aleksey Min, Matthias Scherer, Amelie Schischke \* and Rudi Zagst**

Department of Mathematics, Technical University of Munich, Boltzmannstr. 3, 85748 Garching, Germany;
min@tum.de (A.M.); scherer@tum.de (M.S.); zagst@tum.de (R.Z.)

**\*** Correspondence: amelie.schischke@tum.de

check for
updates

**Abstract:** A sound statistical model for recovery rates is required for various applications in quantitative risk management, with the computation of capital requirements for loan portfolios as one important example. We compare different models for predicting the recovery rate on borrower level including linear and quantile regressions, decision trees, neural networks, and mixture regression models. We fit and apply these models on the worldwide largest loss and recovery data set for commercial loans provided by GCD, where we focus on small- and medium-sized entities in the US. Additionally, we include macroeconomic information via a predictive Crisis Indicator or Crisis Probability indicating whether economic downturn scenarios are expected within the time of resolution. The horserace is won by the mixture regression model which regresses the densities as well as the probabilities that an observation belongs to a certain component.

**Keywords:** decision tree; loss given default; mixture model; neural network; predictive crisis indicator

## 1. Introduction

Additional capital requirements and an increased awareness of the importance of credit risk modeling are consequences of the financial crisis of 2007. Capital requirements, like the internal ratings-based approach of Basel II, allow financial institutions to estimate their credit risk using internal models. The main determinants of credit risk are the probability of default (PD), the exposure at default (EAD), and the loss given default (LGD); the latter is linked to the recovery rate (RR) via $RR = 1 - LGD$. We focus on the modeling of the recovery rate and compare different methods to estimate a firm-specific one.

According to §297 of [1], LGD has to be measured as "loss given default as a percentage of the EAD". However, there exist several methods to calculate the LGD (resp. RR), namely the market LGD, the implied market LGD, and the workout LGD (see [2]). For loan data, the appropriate definition is the workout RR, which are the revenues (R) that financial institutions can collect reduced by all administration costs (A) during the resolution period in case of a default, divided by the outstanding amount at default (EAD). In [3], the $RR$ for a defaulted loan with exposure at default $EAD$ is defined as:

$$RR = \frac{R - A}{EAD} = \frac{\sum \text{Collections} - \sum \text{Admin Fees}}{\text{Outstanding Balance at Default}}.$$

where $A$ denotes the administration costs and $R$ the recovered amount. With this definition, it is actually possible that the RR can take values greater than one or smaller than zero. In [4], an example for a RR smaller than zero principal advances is mentioned. On the contrary, in cases of penalty fees, additional interest, and recovered principal advances, the RR can attain a value greater than one. Both are frequently observed in our data.

Our objective is to compare different methods to model the recovery rate, namely linear regression, quantile regression, decision trees with linear/quantile regression, neural networks, and mixture regression models. Thereby, we investigate how information on an economic crisis affects these models.

We base this study on the LGD&EAD platform provided by Global Credit Data (GCD) which contains information about defaulted loans; see Section 4.1 for more details. Moreover, we aggregate the information on loan level to borrower level and focus on small- and medium-sized entities (SME) in the US. Inspired by the findings presented in [5], in which it was stated that the macroeconomic behavior during the resolution time has an influence on the recovery rate, we include a predictive Crisis Indicator, which predicts whether a crisis might occur during the time of resolution, in our models.

This paper is structured as follows: In Section 2, we survey the literature for modeling the RR. Subsequently, we provide a theoretical background of the techniques used in this study in Section 3. Thereafter, in Section 4, the structure of the database is presented. The results of the models are shown in Section 5. Section 6 summarizes the results and we discuss possibilities for further research on the RR.

## 2. Literature Review

In the literature, several models are suggested to estimate the RR. We give a short overview with focus on regression models, decision trees, neural networks, and mixture models.

According to the authors of [6,7], classical linear regression models are the most popular and most straightforward techniques to estimate the RR. However, the authors of [8,9] mention as a drawback that in reality, RRs are bounded and not normally distributed. Nevertheless, the linear regression model outperforms the Tobit model and the decision tree model for UK credit card accounts in [10].

Many authors have adapted regression models to the situation of RRs. The inverse Gaussian (IG) regression transforms the RR by an inverse Gaussian distribution function from the interval $(0, 1)$ to the real line. The authors of [9] compare this to the inverse Gaussian regression with beta transformation, which is also used by the authors of [7,11,12], where the assumption of beta distributed LGDs is postulated and subsequently, the inverse Gaussian distribution is applied. Linear regression aims at predicting the mean, whereas a quantile regression can analyze the influence of covariates on the entire distribution. The authors of [13] emphasize that quantile regression might hence be better suited for downturn scenarios. Moreover, the authors of [14] use SME data from the biggest Polish banks to compare a linear regression, quantile regression, and the standard quantile regression forests with the weighted quantile regression forests and conclude that the weighted quantile regression forests outperform the other methods.

In order to model the concentration of RRs at the boundaries $\{0, 1\}$, the authors of [10] propose a decision tree model which is also used by the authors of [7]. A logistic regression model decides whether the RR takes the value 0 or 1. Subsequently, an ordinary least squares method is used inside $(0, 1)$. Similarly, the authors of [12] use a logistic regression to determine whether the RR takes the boundary values and different parametric as well as non-parametric models to explain the RR inside $(0, 1)$, but the single application of the non-parametric models, especially neural networks and the least squares support vector machines, outperforms the combinations.

Beside the authors of [12], several other researchers also studied non-parametric models: In [6,9], neural networks outperform the fractional response regression resp. linear regression, inverse Gaussian regression, inverse Gaussian regression with beta transformation, and the fractional response regression. However, [9] mention as a drawback that neural networks are a "black box" because there is no straightforward method to interpret the relationship between the independent and dependent variables.

Another type of models, which are considered in different ways to predict the RR, are finite mixture models. The authors of [13] use a normal mixture distribution with two components for LGDs and find that it performs best with their quantile regression on the GCD subset of US SMEs. However, the authors of [3] propose a two-stage model to apply a beta-mixture model for the RRs in $(0, 1)$.

This two-stage model outperforms the OLS, the OLS with lasso, and the beta regression. In addition, the authors of [15] present zero-and-one inflated mixture models. A three-level multinomial model first decides whether the LGD takes the value 0 or 1 or lies in $(0, 1)$. Subsequently, finite mixture distributions are applied to $(0, 1)$, in which they test different component distributions.

In [16], the transformed RR (by the inverse normal distribution) is approximated as a mixture of Gaussian distributions, where only the probability belonging to a certain component depends on covariates. The authors of [17] extend this mixture model on the Moody's Ultimate Recovery Database by introducing a Markov switching model with two states representing crisis and non-crisis periods to capture cyclical aspects. For each state, there is a mixture model with four components for the transformed RR which enables the determination of the influence of covariates.

Similar to the decision trees, a mixed continuous–discrete model is presented in [18]. In her further work, Calabrese [19] extends her model by introducing a mixture model. The LGD is modeled as a mixture of the expansion and recession distribution where each distribution is represented by the mixed model in [18]. The mixtures represent the credit cycle, i.e., bad or good times.

## 3. Modeling Methods

This section provides a theoretical background of the techniques used in this study. We focus on decision trees, neural networks, and mixture regression models. We refer to [13,20,21] for more information on regression methods, in particular for quantile regression models as well as for model selection techniques.

### 3.1. Rule Based Models—Decision Trees

Since RRs are not normally distributed, a linear regression might not be adequate. As an alternative, the RR can first be transformed and then, on the transformed data, a linear regression can be applied. In the literature, e.g., in [11], a beta transformation is used. The transformed RR is:

$$\text{Transformed RR} = \Phi^{-1}(F_{Beta}(RR, \alpha, \beta)), \tag{1}$$

where $\Phi^{-1}$ is the quantile function of the standard normal distribution and $F_{Beta}(x, \alpha, \beta)$ is the distribution function of the beta distribution with shape parameters $\alpha$ and $\beta$, which have to be estimated. This transformation, however, can only be applied to $RR \in (0, 1)$. As our data set also contains observations with a RR smaller than or equal to zero or greater than or equal to one, we use a rule-based model as displayed in Figure 1. The model is similar to a decision tree approach (see, e.g., [10]), and we use the terminology decision trees in what follows.
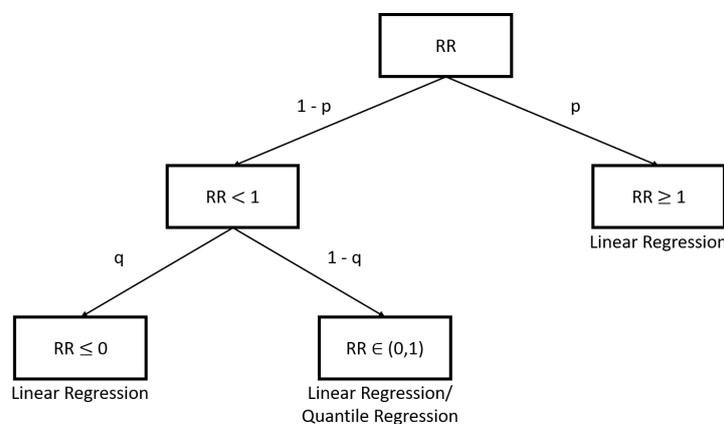


**Figure 1.** Structure of the decision tree model.

Firstly, a logistic regression (or a neural network) determines the probability $p$ that the RR is greater than or equal to 1. Then, a second classification model, i.e., a logistic regression (resp. neural

network), estimates the probability $q$ that the RR takes a value less than or equal to 0, given that it is smaller than 1. We use a linear regression to predict the rates $RR_{\geq 1}$ and $RR_{\leq 0}$. Inside $(0, 1)$, we apply the beta transformation (1) to the RR. Subsequently, a linear regression (or a quantile regression) estimates the rate $RR_{(0,1)}$. If the linear regression had been applied to the raw RR, there would be predicted values outside $(0, 1)$. Therefore, we first apply the beta transformation and, on the transformed RR, we can use the linear regression. In contrast to the linear regression, the estimates of the quantile regression would not exceed the open unit interval. Therefore, we apply this regression type on the raw RR and compare the results. We mention that, according to our results, it is better to apply the quantile regression on the raw $RR \in (0, 1)$. Hence, in the following, the corresponding results are presented. The expected RR is expressed as a weighted average, where the weights are $p$, $(1 - p) \cdot q$, and $(1 - p) \cdot (1 - q)$. Hence, the expected RR is:

$$E[RR] = p \cdot RR_{\geq 1} + (1 - p) \cdot q \cdot RR_{\leq 0} + (1 - p) \cdot (1 - q) \cdot RR_{(0,1)}.$$

*3.2. Neural Networks*

In this section, we present the structure of feedforward neural networks following [22,23]. For a comparison of statistical models and neural networks, please refer to [24].

In a neural network, whose structure is presented in Figure 2, neurons are structured in layers. The neurons are connected by synapses, which are graphs between them and the neurons of the subsequent layers. In order to keep the model simple, we (mostly) consider feedforward neural networks with one hidden layer. The input layer contains all covariates, the so-called input variables $X_1, \ldots, X_p$, which represent the separate neurons. Each numerical attribute has its own neuron. In case of categorical variables, dummy coding as in a linear regression is applied. The output layer has $K$ neurons $O_1, \ldots, O_K$. For regression problems with one response variable as well as for classification problems with two categories, we have $K = 1$. For classification problems with $C$ classes, there are $K = C$ output neurons, each representing one category. The hidden layer with neurons $H_1, \ldots, H_M$ lies in between and cannot be observed directly. A bias can be added to the input and hidden layers as an extra neuron $B_I$ resp. $B_H$.
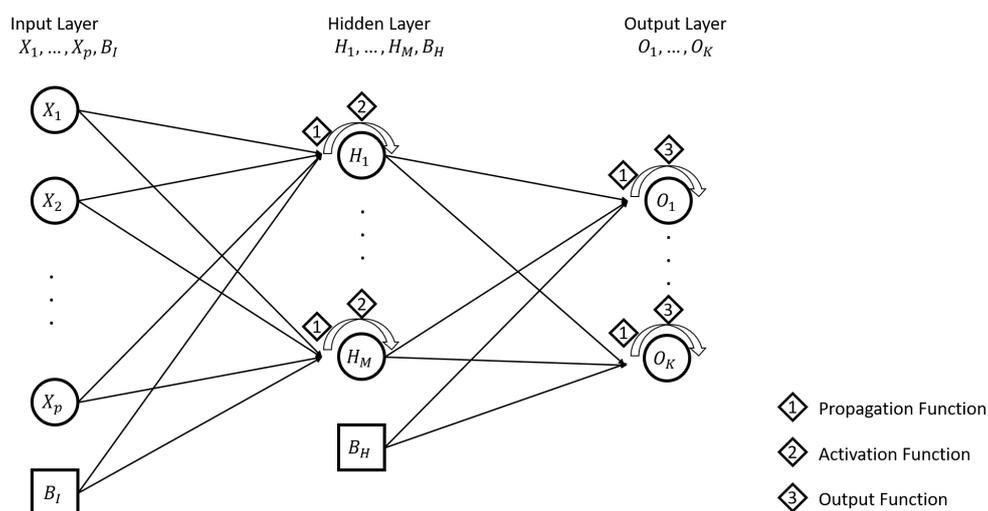


**Figure 2.** Structure of neural networks with one hidden layer.

The propagation function connects the output values of the previous layer $O_{j,\text{previous layer}}$, $j \in$ previous layer such that the result can be used as input $I_{i,\text{current layer}}$ for a neuron $i \in$ current layer in the current layer. We use the weighted sum:

$$I_{i,\text{current layer}} = \sum_{j \in \text{previous layer}} w_{i,j} O_{j,\text{previous layer}}.$$

The activation function $\sigma$ transforms this value $I_{i,\text{current layer}}$ to the output value of the neuron $O_{i,\text{current layer}} = \sigma(I_{i,\text{current layer}})$. For this, we use the sigmoid function:

$$\sigma(x) = \text{sigmoid}(x) = \frac{1}{1 + \exp(-x)}.$$

The propagation function is applied again to receive the input for the output layer. Then, for the output neuron $O_k$, $k = 1, \ldots, K$, in the output layer, we apply a final transformation by the output function $g_k$ instead of the activation function. In case of regression problems, we use as $g_k$ the identity function, whereas we apply the softmax function $g_k(I_k) = \frac{\exp(I_k)}{\sum_{l=1}^{K} \exp(I_l)}$ in case of classification problems.

The weights have to be estimated in the training process. Therefore, we use the backpropagation algorithm in case of neural networks for classification problems as "backpropagation is the most widely used algorithm for supervised learning with multilayered feed-forward networks" according to the authors of [25]. In case of regression problems, we use the extension RPROP+ algorithm of the authors of [25] and refer to their original paper for more information.

### 3.3. Mixture Models

In a linear regression model, we assume that the dependent variable relates to the covariates by a fixed parameter $\beta$ over all observations. This assumption is often too restrictive, calling the need for models in which the regression coefficient can change over different clusters among the observations. One family of models are finite mixture models, which will be presented following [26–29].

In general, a finite mixture regression model with $K$ components has the form:

$$h(y|x, \psi) = \sum_{k=1}^{K} \pi_k f(y|x, \theta_k), \tag{2}$$

where $\pi_k$, $k = 1, \ldots, K$ are the weights with $\pi_k \geq 0$, $\sum_{k=1}^{K} \pi_k = 1$ and $\psi = (\pi_1, \ldots, \pi_K, \theta'_1, \ldots, \theta'_K)$ is the vector of all unknown parameters. $\theta_k$ denotes the component specific parameter vector for the density function $f$. If $f$ is a univariate normal density with component specific mean $\beta'_k x$ and variance $\sigma_k^2$, we get a mixture of standard linear regression models with $\theta_k = (\beta'_k, \sigma_k^2)'$.

The weights $\pi_k$, $k = 1, \ldots, K$ in Equation (2) are usually independent of the covariates. One extension is the concomitant variable model by the authors of [30], which assumes that the weights depend on some variables, the so-called concomitant variables denoted by $c$. Then, the mixture model can be written as:

$$h(y|x, \psi) = \sum_{k=1}^{K} \pi_k(c, \alpha) f(y|x, \theta_k), \tag{3}$$

where $\alpha$ denotes the parameter vector of the concomitant variable model and $\psi$ contains all parameters including $\alpha$. The remaining arguments are defined as in Equation (2) and the weights have to satisfy the conditions $\pi_k(c, \alpha) > 0$ and $\sum_{k=1}^{K} \pi_k(c, \alpha) = 1$, $k = 1, \ldots, K$. Similar to the authors of [30], we assume a multinomial logit model for the weights $\pi_k$, which can be written as:

$$\pi_k(c, \alpha) = \frac{\exp(c'\alpha_k)}{\sum_{u=1}^{K} \exp(c'\alpha_u)}, \tag{4}$$

for all $k = 1, \ldots, K$ and with $\alpha = (\alpha'_k)'_{k=1,\ldots,K}$ and $\alpha_1 \equiv 0$.

For parameter estimation, we write the log-likelihood function of a sample of $n$ observations $(x_1, y_1), \ldots, (x_n, y_n)$ as:

$$\log L = \sum_{i=1}^{n} \log h(y_i|x_i, \psi) = \sum_{i=1}^{n} \log \left( \sum_{k=1}^{K} \pi_k(c, \alpha) f(y_i|x_i, \theta_k) \right).$$

Since the membership to the components is unknown, this likelihood function cannot be computed directly. For maximum likelihood estimation of mixture models with concomitant variables, the authors of [27] outline an iterative expectation–maximation (EM) algorithm introduced by the authors of [31] and implement it in the R-package `flexmix`.

## 4. Data

### 4.1. The Global Credit Data (GCD) Database

In line with the authors of [13], for our empirical analysis, we use a data set of US-based small- and medium-sized entities (SME) from Global Credit Data (GCD). GCD is a Dutch-based, not-for-profit registered association whose owners are more than 50 member banks across the world. The objective of GCD is to be a credit risk data pooling initiative to support the member banks by their internal credit risk models inter alia for the advanced internal ratings-based approach of Basel II. We use the LGD&EAD platform, which is the largest loss and recovery data set for commercial loans worldwide and contains data relating to credit defaults from 1998 up until the end of 2016. This time period encompasses more than one full economic cycle as required by §472 in [1]. Table A1 in the Appendix A gives an overview of all variables used.

We adjust the data following [32]. Firstly, the exposure at default has to be strictly greater than zero, as the focus of this study lies on real losses. Second, we only consider loans where EAD + Principal Advance + Financial Claim $\geq$ €5000, such that very small exposures are excluded. Third, the default date lies in the interval [January 2002, December 2015]. We exclude cases before the year 2002 due to modified banking regulations. As the cases after 2015 might still be unresolved, we exclude them as well. Fourth, to exclude all facilities that are not fully resolved or exhibit unreasonable cash flows, the following rule is applied according to [32]: If the total sum of all reported cash flows (including charge-offs and waivers) divided by the outstanding amount at default is smaller than 90% or greater than 105%, the facility is not considered. Fifth, only cases with resolved default status are of interest. Finally, the RR lies in the interval $[-0.5, 1.5]$. All observations with smaller or greater RR are excluded to avoid outliers.

Furthermore, we split the data into three groups: training, validation, and test set. The training set (in regression problems, the so-called in-sample set) contains 80% of the data according to [29] and is used to estimate the models. In order to get an impression of how well a model can create new predictions, the trained models are applied on the test set, which is also called the out-of-sample set. These data are not used in the estimation of the model; therefore, these results are reliable and can be compared. Some models need hyperparameters—for example, the number of hidden neurons in a neural network. Since the training data are already used and the test data should remain independent of the modeling process, we use a third data set, the validation set, to fit the hyperparameters. The test set and the validation set both contain 10% of the data.

The histogram of the RR is presented in Figure 3 and shows a high concentration at full recovery. Furthermore, there are two additional peaks near 0 and 0.5. In the literature, the RR has frequently been modeled using a bimodal structure—for example, in [6,9,16,33]. Similar to our data, the authors of [3] use a trimodal distribution.

**Histogram of the Recovery Rate**
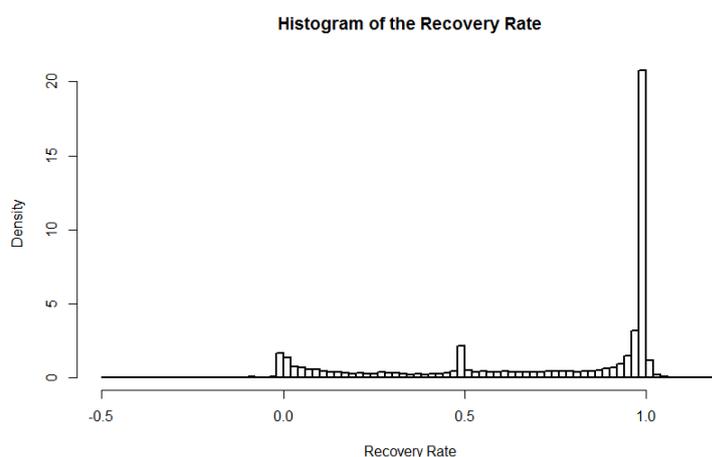


**Figure 3.** Histogram of the recovery rate in our data (small- and medium-sized entities (SME), US-based).

### 4.2. Predictive Crisis Indicator

Some studies—for example, [18,32]—find that the recovery rate tends to be lower during economic downturns. The authors of [5] observe that the macroeconomic behavior during the resolution time has an influence on the recovery. Therefore, we use a predictive Crisis Indicator, which indicates whether a crisis might occur in the next 18 months (the average resolution time).

To model the predictive Crisis Indicator, we first calculate a daily Crisis Indicator using a modified version of the algorithm of [34], where we use two-year highs instead of half-year highs. The algorithm of [34] can be applied to any stock index, but for the focus on recovery rates of SMEs in the US, we chose the S&P500. With the algorithm from [34], a daily crisis indicator is determined. To get a monthly aggregated crisis indicator, we apply the following decision rule: If at least two days within a month are indicated as crisis, the month in total is considered as crisis.

In the next step, a predictive Crisis Indicator needs to be built; therefore, we follow an approach tested by [35]. For every month $m$, we consider the period of the next 18 months $[m+1, \ldots, m+1+18]$. If there is at least one month in crisis, the predictive Crisis Indicator for $m$ is set to 1 (indicating a crisis).

Up to this point, the calculations are made on historical data and the predictive Crisis Indicator can only be obtained once the data for the next 18 months are available. Since the goal of this study is to predict the RR at the date of default, the required information is not yet available. Therefore, the predictive Crisis Indicator has to be modeled. For this, we set up a logistic regression model with macroeconomic data and Table 1 shows the included attributes and their impact.

**Table 1.** Estimated logistic regression model for the Crisis Indicator.

| Variable | Positive (+) or Negative (−) Impact | Description |
|---|---|---|
| (Intercept) | + | |
| Implied Volatility | + | Implied Volatility |
| (TEDRATE)$^2$ | + | squared TED Spread |
| TCU | − | Capacity Utilization Rate for Total Industry |
| FEDFUNDS | + | Effective Federal Funds Rate |
| FEDFUNDS: (TEDRATE)$^2$ | − | Interaction between FEDFUNDS & squared (TEDRATE) |
| T10Y3MM | + | Spread between 10-Year treasury Constant Maturity and 3-Month Treasury Constant Maturity |
| (OECD_6NME)$^2$ | + | squared Composite Leading Indicator given by the OECD |
| log(UNRATE) | − | Civilian Unemployment Rate |
| OECD_6NME | − | Composite Leading Indicator given by the OECD |

In this paper, the Crisis Indicator is used in different ways:

(nC)　We do not include the crisis information at all.

(CP)　The predicted Crisis Probability calculated from the logistic regression model is included as a covariate.

(CI)　The Crisis Indicator is included as a covariate.

(sC)　We split the data into crisis and non-crisis data sets and train the models on each subset.

## 5. Empirical Results

The main focus of this study lies on mixture regression models. Therefore, we first briefly present the linear and quantile regressions followed by the decision trees as well as the neural networks. Subsequently, we discuss the mixture regression model in detail. Thereafter, we compare all models and conclude with practical consequences.

In order to rank the models, we use the mean squared error (MSE) measure of fit defined as:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$

where $y_i, i = 1, \ldots, n$ are the observed RRs and $\hat{y}_i$ are the estimated RRs. This measure is also used, e.g., in [3,11,18].

### 5.1. Regression Models

First, we consider the results of the regression models. We apply stepwise selection for model selection in the different regression problems based onthe Bayesian information criterion (BIC) of [36], as it penalizes model complexity on a larger scale compared to the Akaike information criterion (AIC) of [37]. In the following, we will use this model selection criterion in every regression problem. We concentrate on the linear regression model including the Crisis Indicator as well as the models trained on the crisis and non-crisis subsets, since the results of these models performed best (see Section 5.5). Regarding the included covariates within the linear models, which are presented in Table 2, we recognize that in case of crisis, the RR is only determined by the information whether a guarantee or collateral is given and the size of EAD. Moreover, in the crisis case, the Collateral Indicator has an impact on the RR, whereas the primary industry code as well as the Utilization Rate only have an impact in the non-crisis case. However, the linear regression model with the Crisis Indicator includes all these attributes and, additionally, the variable nature of default.

**Table 2.** Included attributes in the linear regression models (LR) on the crisis/non-crisis subset resp. the entire subset including the Crisis Indicator and in the quantile regression model (QR) including the Crisis Indicator.

| Variable | LR on Crisis Subset | LR on Non-Crisis Subset | LR incl. Crisis Indicator | QR incl. Crisis Indicator |
|---|---|---|---|---|
| Crisis Indicator | | | x | x |
| Country of Business | | | | x |
| Leveraged Finance Indicator | | | | x |
| Operating Company Indicator | | | | x |
| Primary Industry Code | | x | x | x |
| Nature of Default | | | x | x |
| Collateral Rank of Security | | | | x |
| Guarantor Rating Moody's | | | | x |
| Guarantee Indicator | x | x | x | x |
| Collateral Indicator | x | | x | x |
| log(EAD) | x | x | x | x |
| Utilization Rate | | x | x | x |

In the case of the quantile regression model, we regress the median in order to compare the results of the quantile regression to the results of the linear regression. The quantile regression

including the Crisis Indicator outperforms the remaining models considering MSE (see Table 6). Therefore, we have a look at the included variables of this model and recognize that the model selection results in more attributes for the quantile regression as in the cases of the linear regression models. In particular, the variables Country of Business, Leveraged Finance Indicator, Operating Company Indicator, Collateral Rank of Security, and Guarantor Rating Moody's are only included in the quantile regression model, whereas the remaining covariates are part of both models.

*5.2. Decision Tree*

We use a decision tree approach in order to apply the beta transformation on the RR. Similar to the regression models in Section 5.1, we use stepwise selection with BIC for model selection.

Besides the logistic regression, a neural network can be applied for the categorization problems. Therefore, we use the function `nnet` of the eponymous R-package. As input, all available attributes are used. The network is trained by the backpropagation algorithm and only has one hidden layer for simplicity. Furthermore, the optimal number of hidden neurons is tested. Therefore, the network is trained on the training data with different numbers of hidden neurons from 1 to 10. The prediction error of the validation set is the selection criteria and results for both classification problems, whether the RR is greater than or equal to 1 and whether the RR is smaller than or equal to 0, in one hidden neuron.In general, a neural network with one hidden neuron and the sigmoid function as activation function equals a logistic regression model. Since the estimation method is different (backpropagation algorithm in case of a neural network and maximum likelihood estimation in case of a logistic regression), the parameters of the two models can be different.

We first report the trees with the quantile regression for the median applied on the raw RR in the open unit interval because they outperform those on the beta-transformed RR. One reason for this might be that the RR in our data set is trimodal and the beta distribution might not fit well.

Table 3 displays which covariates are included in the different regression problems. We focus in the unit interval on the linear regression model without crisis information and the quantile regression including the Crisis Probability, as the results of the decision trees including these models outperform the other decision trees. Similar to the regression problems above, the quantile regression model including the Crisis Probability contains more variables than the linear regression model. Furthermore, the Crisis Indicator affects only the logistic regression to decide whether the RR attains a value smaller than or equal to zero.

**Table 3.** Included attributes in the regression models (linear regressions (LR), quantile regressions (QR), and logistic regressions (Log. Reg.) of the decision trees.

| Variable | LR without $RR \in (0,1)$ Crisis info for $RR \in (0,1)$ | QR incl. $RR \in (0,1)$ Crisis Probability for $RR \in (0,1)$ | LR for $RR \geq 1$ | LR for $RR \leq 0$ | Log. Reg. $RR \geq 1$ | Log. Reg. $RR \leq 0$ |
|---|---|---|---|---|---|---|
| Crisis Indicator | | | | | | x |
| Crisis Probability | | x | | | | |
| Country of Business | x | x | | | x | |
| Public Private Indicator | | | x | x | x | |
| Leveraged Finance Indicator | x | x | | | | |
| Operating Company Indicator | x | x | | | | |
| Primary Industry Code | | x | | | | |
| Nature of Default | | x | | x | x | x |
| Lender Issued Amount | | | | x | | |
| Collateral Rank of Security | x | x | | | | |
| Guarantor Rating Moody's | | x | | | | |
| Guarantee Indicator | x | x | | | | x |
| Collateral Indicator | x | x | x | | | |
| Entity Sales | | | x | | | |
| log(Number of Loans) | | | | | x | |
| log(EAD) | x | x | | | | x |
| Utilization Rate | x | x | x | | x | |

*5.3. Neural Network*

Another possibility to model the RR are neural networks whose results are shown in the following. We begin with the description of the predetermined model parameters and present the results in Table 6.

In order to train neural networks for regression problems, we use the function `neuralnet` of the R-package `neuralnet`, which applies the RPROP+ algorithm. Therefore, we set the multiplication factors for the upper and lower learning rate to $\eta^- = 0.5$ and $\eta^+ = 1.2$, the parameter threshold to 0.01, and the maximum number of iterations to $1 \times 10^7$. We use the sigmoid activation function, the identity as output function, and the sum of squared errors as error function.

For reasons of simplicity, all neural networks contain one hidden layer and the optimal number of hidden neurons is determined by minimizing the MSE on the validation set. We also computed neural networks with two hidden layers; however, these networks performed not significantly better (differences only in the third or fourth digit of MSE) than the networks with one hidden layer. Therefore, we decided to focus on networks with one hidden layer. For a pre-selection and in order to reduce the number of input neurons, we include the input variables log-transformed EAD, Guarantee Indicator, Collateral Indicator, and Primary Industry Code, similar to [13], as well as Utilization Rate and Number of Loans and the Crisis Indicator.) For the categorical covariates, dummy variables are created just like for the regression problems. According to [38], we scale the metric data to the unit interval $[0, 1]$ as it is not normally distributed.

### 5.4. Mixture Regression Models

In this subsection, we present the results of the mixture regression models from Section 3.3. At first, we apply the model selection to identify relevant covariates. Subsequently, the results of the mixture models with and without concomitant variables are shown.

### 5.4.1. Model Parameters

Our motivation to investigate mixture regression models stems from the observation of multiple modes (see Figure 3). To determine the number of mixture components, we fit simple mixture regression models with constant parameters for $K = 1, \ldots, 6$ components to the in-sample data and select the best model based on the BIC. The resulting model consists of five mixed normal densities, but three of them have their expectation in a narrow interval of length 0.043 around one. Therefore, we decided to choose a more parsimonious mixture regression model with three components. This decision is also supported by the trimodality of RR in our data set.

We use the package `flexmix` in R to fit our mixture regression models. For a pre-selection and in order to reduce the overall number of covariates, we use the input variables from [13], the authors of which base their study on loan data of SMEs in the US provided by GCD. We use our crisis information instead of macroeconomic data and focus our analysis on entity level, hence we can not use all of the attributes from [13]. In conclusion, the resulting variables are log-transformed EAD, Guarantee Indicator, Collateral Indicator, Primary Industry Code, and the Crisis Indicator resp. Probability. Subsequent to this pre-selection, all possible combinations of variables are formed and it is tested whether the EM-algorithm converges. We compare the mixture regression models with the BIC and the best model contains the covariates Collateral Indicator, log-transformed EAD, and Crisis Probability. For simplicity reasons, this combination of variables is also used as concomitant variables in the following. We notice that it is better to use the Crisis Probability than the Crisis Indicator in this method.

The package `flexmix` provides information about the standard error as well as z- and p-value for every coefficient in every component. In case of a negative entry in the diagonal of the variance-covariance matrix, the standard error can not be computed. This is partially the case for the coefficient of the log-transformed attribute EAD.

Therefore, we exclude this attribute for the regression problems, but it is still part of the multinomial model. In order to distinguish the different models, we refer to the model including the EAD as covariate by the name "MwEAD" and denote the model without the EAD as "Mw/oEAD". If, additionally to the components, the probabilities of belonging to the components are regressed,

we denote the models "CMwEAD" and "CMw/oEAD", since the attributes included in the multinomial models are called concomitant variables.

### 5.4.2. Model Description

Table 4 displays the results of the estimated models. Firstly, we consider model MwEAD. The first component of MwEAD is mainly determined by the intercept near one, since the remaining coefficients are vanishing. Furthermore, the intercept at 0.5 as well as the attribute Collateral Indicator have the most impact on the second cluster. In contrast to the other clusters, the characteristic Yes of the Collateral Indicator has a negative impact on the third cluster. Moreover, the influence of the characteristic No of the Collateral Indicator as well as the Crisis Probability is negative. In addition, the EAD has a slightly positive impact. The third component has the highest fluctuations (represented in a sigma of 0.308).

**Table 4.** Summary of the estimated mixture regression models MwEAD (inclusion of the attribute EAD), Mw/oEAD (exclusion of the attribute EAD), and the estimated concomitant variable models CMwEAD (inclusion of the attribute EAD) and CMw/oEAD (exclusion of the attribute EAD).

| **Comp. 1** | | | | |
| --- | --- | --- | --- | --- |
| | **MwEAD** | **Mw/oEAD** | **CMwEAD** | **CMw/oEAD** |
| (Intercept) | 0.9967 | 0.9920 | 0.9796 | 0.8051 |
| Collateral Indicator: No | −0.0045 | −0.0044 | −0.0111 | 0.0151 |
| Collateral Indicator: Yes | 0.0030 | 0.0014 | −0.0057 | −0.0285 |
| log(EAD) | −0.0003 | - | −0.0004 | - |
| Crisis Probability | −0.0008 | −0.0012 | −0.0055 | 0.0006 |
| Sigma | 0.0052 | 0.0077 | 0.0287 | 0.1818 |
| **Comp. 2** | | | | |
| | **MwEAD** | **Mw/oEAD** | **CMwEAD** | **CMw/oEAD** |
| (Intercept) | 0.5158 | −0.0143 | 0.9978 | 0.9921 |
| Collateral Indicator: No | −0.0087 | 0.0280 | −0.0057 | −0.0045 |
| Collateral Indicator: Yes | 0.4687 | 0.0362 | 0.0015 | 0.0014 |
| log(EAD) | −0.0007 | - | -0.0003 | - |
| Crisis Probability | −0.0089 | 0.0206 | −0.0007 | −0.0008 |
| Sigma | 0.0195 | 0.0366 | 0.0053 | 0.0074 |
| **Comp. 3** | | | | |
| | **MwEAD** | **Mw/oEAD** | **CMwEAD** | **CMw/oEAD** |
| (Intercept) | −0.3386 | 0.6348 | −0.3593 | 0.3214 |
| Collateral Indicator: No | −0.2457 | −0.1046 | −0.1549 | −0.0410 |
| Collateral Indicator: Yes | −0.2637 | 0.0465 | −0.2147 | −0.2223 |
| log(EAD) | 0.0814 | - | 0.0763 | - |
| Crisis Probability | −0.0070 | -0.0505 | 0.0172 | 0.0720 |
| Sigma | 0.3077 | 0.2702 | 0.2858 | 0.1744 |

Having a look at Mw/oEAD, we notice that the first component is mainly influenced by the intercept near one and the attributes have little impact. In comparison to the control group Unknown, the categories Yes and No of the Collateral Indicator have a positive impact on the second component. As the parameter of Yes of the Collateral Indicator is higher than the coefficient for No, we would expect higher values for entities having collateral. It is counterintuitive that this component increases its value if the Crisis Probability increases. The highest values of the third component are expected in the case of collateral, whereas the lowest values will be attained when there is no collateral. In addition, the value of this cluster will be higher if the Crisis Probability is small. The attributes have the highest impact on the third cluster due to their higher absolute values. Sigma with a value of 0.270 underlines this finding, as it is higher than the sigma of the first and second cluster.

The first as well as the second component of the mixture regression model CMwEAD are mainly determined by the intercept near one, as the coefficients of the covariates are near zero and influence them little. Sigma of the third component is the highest, indicating higher fluctuations. The characteristics of the Collateral Indicator influence the third component negative due to their negative parameters. Moreover, the log-transformed EAD has a positive coefficient, which indicates that the value of the third cluster correlates to the EAD. In addition, a higher Crisis Probability results in a higher value for this component which is counterintuitive.

In model CMwEAD, we regress the probabilities that an observation belongs to a certain component. In this study, a multinomial logit model is assumed for the weights $\pi_k$, $k = 1, \ldots, K$ as depicted in Equation (4). One assumption of this model is $\alpha_1 \equiv 0$. Therefore, only the parameters of the second and third component are given in Table 5.

**Table 5.** Summary of the estimated concomitant variable models CMwEAD (inclusion of the attribute EAD) and CMw/oEAD (exclusion of the attribute EAD).

| Comp. 2 | | |
| --- | --- | --- |
| | **CMwEAD** | **CMw/oEAD** |
| (Intercept) | 1.5822 | 3.6216 |
| Collateral Indicator: No | 1.1220 | 0.8513 |
| Collateral Indicator: Yes | −0.1528 | −0.0661 |
| log(EAD) | −0.0345 | −0.2217 |
| Crisis Probability | −0.4787 | −0.5281 |
| **Comp. 3** | | |
| | **CMwEAD** | **CMw/oEAD** |
| (Intercept) | 2.2059 | 7.0839 |
| Collateral Indicator: No | 1.4026 | 2.0530 |
| Collateral Indicator: Yes | −0.1795 | 0.2222 |
| log(EAD) | −0.1054 | −0.6220 |
| Crisis Probability | 0.2783 | 0.3027 |

The probability of belonging to the second component increases if no collateral is available and decreases if collateral is given. Moreover, a borrower with lower EAD (resp. a lower Crisis Probability) is expected to have a higher probability of belonging to the second cluster. Furthermore, the probability that an entity belongs to the third cluster is expected to be lower in the case of collateral and higher in the case of no collateral. The log-transformed EAD has again a negative influence, whereas the Crisis Probability has a positive one.

In model CMw/oEAD, we would expect that the value of the first component is higher in case of no collateral than in case of collateral. In addition, a higher Crisis Probability will lead to higher values of the first component. This cluster has the highest variation which is displayed in the high value of sigma of 0.182. The covariates have little impact on the second component, which is mainly determined by its intercept near one. Moreover, the value of the third cluster is lower for an observation with a collateral than for one without any collateral. Furthermore, the third component is expected to attain a higher value for a higher Crisis Probability.

The probability that an observation belongs to the second cluster increases if it has no collateral. However, having a collateral decreases the probability. In addition, the higher the log-transformed EAD or the higher the Crisis Probability, the lower the probability that an observation belongs to the second cluster. Furthermore, the probability that an observation belongs to the third cluster reaches a maximum if there is no collateral. Moreover, we would expect a lower probability that an entity belongs to the third cluster if the EAD is high or the Crisis Probability is low.

*5.5. Comparison of All Models Based on MSE*

Table 6 shows the in-sample as well as the out-of-sample results for all models including the linear regression model with different assumptions (nC), (CI), (CP), (sC) on the crisis information. For the linear regression, MSE prefers in-sample the model including the Crisis Indicator and out-of-sample to separate the data into crisis and non-crisis subsets. Since the models on the split data give more insights in the determinants of the RR in crisis and non-crisis case, this approach might be preferred, as the goodness-of-fit of the models is similar.

**Table 6.** In-sample and out-of-sample MSE for the estimated linear regression (LR), quantile regression (QR) models, decision trees (DT) with linear or quantile regression in the unit interval (LR/QR), and logistic regression or neural network for the classification problems (LogReg or NN), neural networks (NN), and mixture regression models (MwEAD, Mw/oEAD, CMwEAD, CMw/oEAD) for the four different cases nC (no crisis information), CI (inclusion of Crisis Indicator), CP (inclusion of Crisis Probability), sC (split data into crisis and non-crisis data set).

| | In-Sample | | | | Out-Of-Sample | | | |
|---|---|---|---|---|---|---|---|---|
| | **(nC)** | **(CI)** | **(CP)** | **(sC)** | **(nC)** | **(CI)** | **(CP)** | **(sC)** |
| LR | 0.1095 | 0.1085 | 0.1086 | 0.1087 | 0.1095 | 0.1089 | 0.1089 | 0.1082 |
| QR | 0.1382 | 0.1310 | 0.1312 | 0.1377 | 0.1418 | 0.1352 | 0.1353 | 0.1404 |
| DT LR LogReg | 0.1432 | 0.1435 | 0.1434 | 0.1457 | 0.1458 | 0.1465 | 0.1464 | 0.1475 |
| DT LR NN | 0.1395 | 0.1407 | 0.1407 | 0.1431 | 0.1417 | 0.1432 | 0.1432 | 0.1441 |
| DT QR LogReg | 0.1546 | 0.1504 | 0.1504 | 0.1590 | 0.1575 | 0.1542 | 0.1540 | 0.1624 |
| DT QR NN | 0.1513 | 0.1476 | 0.1476 | 0.1569 | 0.1541 | 0.1513 | 0.1511 | 0.1596 |
| NN | 0.0471 | 0.0467 | 0.0454 | 0.0995 | 0.1032 | 0.1011 | 0.0981 | 0.1010 |
| MwEAD | | | 0.0375 | | | | 0.0379 | |
| Mw/oEAD | | | 0.0257 | | | | 0.0268 | |
| CMwEAD | | | 0.0323 | | | | 0.0326 | |
| CMw/oEAD | | | 0.0101 | | | | 0.0107 | |

Having a look at the quantile regression, the model including the Crisis Indicator outperforms the remaining models considering MSE. Moreover, the linear regression models outperform the quantile regressions. This fact might be explained by the different optimization problems. The estimation method of the linear regression minimizes the least squares error, whereas the quantile regression for the median minimizes the mean absolute error. However, the quantile regression models give more insights into the structure of the distribution, since different quantiles can be modeled. We refer to [13] who calculate further quantile regressions for several quantiles.

Comparing the decision trees by MSE, the models with the linear regression in $(0,1)$ outperform those with the quantile regression. Moreover, it is preferable to not include any information about a crisis for the decision tree with the linear regression in the unit interval, whereas the decision tree with quantile regression including the Crisis Probability outperforms the other decision trees with quantile regression. Additionally, the trees with the neural networks for classification result in lower MSEs than the models with the logistic regressions. This finding might be explained by the slightly lower prediction error of the neural network for the classification whether the RR attains values smaller than or equal to 0. However, the differences in the MSE are marginal and since the logistic regression gives some more insight in the determinants of the RR, it might be preferable to use them. We compare the results of the decision tree approach with the regressions on the entire data set and recognize that the models on the entire data result in lower MSEs than the decision trees.

Regarding the results of the neural networks, we notice that the network including the Crisis Probability outperforms the other neural networks in-sample as well as out-of-sample. Furthermore, we recognize that the MSE is in-sample small, whereas the results are out-of-sample similar to the MSE

of the logistic regression models. One reason for the difference in the MSE between the in-sample and out-of-sample subset might be overfitting.

Finally, we consider the results of the mixture regression models. The models excluding EAD as covariate are superior to the mixture regression models including the covariate EAD. In addition, the models which regress the densities as well as the probabilities outperform the mixture models with fixed probabilities. In conclusion, model CMw/oEAD is the best model.

In-sample as well as out-of-sample, the mixture regression models outperform the regressions as well as the neural networks. One reason for this might be that the mixture regression model can display the different modes better than the other models.

A quantitative comparison of our results with other studies is difficult for several reasons (different databases, considered time periods or variables). Nevertheless, let us briefly scrutinize the findings of related studies. Similar to the authors of [13], we focus our analysis on recovery data of small- and medium-sized entities in the US, provided by Global Credit Data. In addition, we also include the variables log-transformed EAD, Guarantee Indicator, Collateral Indicator and Primary Industry Code. Moreover, we consider further attributes, e.g., the Utilization Rate. Different from [7,13], we aggregate macroeconomic information within a Crisis Indicator resp. Crisis Probability indicating whether economic downturn scenarios are expected within the time of resolution. Our finding is in line with the literature. In [3], the beta-mixture model outperforms linear regression models. In addition, the mixture model in [16] leads to better results out-of-sample as regression models or a non-parametric regression tree. Moreover, the neural networks outperform the regression models in our analysis, similar to the findings in [6,9].

*5.6. Practical Consequences from the Best Models*

In the following, we compare the three best models: The mixture regression model CMw/oEAD, the neural network including the Crisis Probability and the linear regression model with separate subsets for crisis and non-crisis.

We investigate the difference $d_i$ between the predicted RRs $\hat{R}_i$ and the observed RRs $R_i^{obs.}$:

$$d_i = \hat{R}_i - R_i^{obs.},$$

for $i = 1, \ldots, n$ where $n$ is the number of observations. From the risk-managers point of view, a situation in which the RR is conservatively underestimated is favorable compared to a situation in which the RR is overestimated.

Therefore, we are interested in the number of observations where the difference between the predicted RR and the observed RR exceeds a certain threshold $\theta \in \{0.1, \ldots, 0.9\}$ proportional to the overall number of observations:

$$\frac{\#\{d_i > \theta\}}{n}.$$

The results are presented in Table 7 for the in-sample and in Table 8 for the out-of-sample data.

**Table 7.** In-sample results for the difference between the estimated and observed RR.

| | $\frac{\#\{d_i>0.1\}}{n}$ | $\frac{\#\{d_i>0.2\}}{n}$ | $\frac{\#\{d_i>0.3\}}{n}$ | $\frac{\#\{d_i>0.4\}}{n}$ | $\frac{\#\{d_i>0.5\}}{n}$ | $\frac{\#\{d_i>0.6\}}{n}$ | $\frac{\#\{d_i>0.7\}}{n}$ | $\frac{\#\{d_i>0.8\}}{n}$ | $\frac{\#\{d_i>0.9\}}{n}$ |
|---|---|---|---|---|---|---|---|---|---|
| LR | 0.2975 | 0.2503 | 0.2117 | 0.1740 | 0.1310 | 0.0833 | 0.0316 | 0.0108 | 0.0013 |
| NN | 0.2859 | 0.2447 | 0.1960 | 0.1467 | 0.0937 | 0.0575 | 0.0267 | 0.0095 | 0.0015 |
| CMw/oEAD | 0.1379 | 0.0386 | 0.0085 | 0.0009 | 0.0001 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |

**Table 8.** Out-of-sample results for the difference between the estimated and observed RR.

|  | $\frac{\#\{d_i > 0.1\}}{n}$ | $\frac{\#\{d_i > 0.2\}}{n}$ | $\frac{\#\{d_i > 0.3\}}{n}$ | $\frac{\#\{d_i > 0.4\}}{n}$ | $\frac{\#\{d_i > 0.5\}}{n}$ | $\frac{\#\{d_i > 0.6\}}{n}$ | $\frac{\#\{d_i > 0.7\}}{n}$ | $\frac{\#\{d_i > 0.8\}}{n}$ | $\frac{\#\{d_i > 0.9\}}{n}$ |
|---|---|---|---|---|---|---|---|---|---|
| LR | 0.2999 | 0.2613 | 0.2263 | 0.1858 | 0.1288 | 0.0929 | 0.0359 | 0.0110 | 0.0000 |
| NN | 0.3008 | 0.2613 | 0.2061 | 0.1527 | 0.0938 | 0.0607 | 0.0304 | 0.0092 | 0.0009 |
| CMw/oEAD | 0.1398 | 0.0414 | 0.0101 | 0.0009 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |

In-sample, the mixture regression model CMw/oEAD overestimates the true RR by more than $\theta = 0.1$ in 14% of all cases, whereas the linear regression model as well as the neural network overestimates the RR even in 29% of all observations. Having a look at $\theta = 0.2$, we recognize that the mixture regression model CMw/oEAD only overestimates the observed RR in 3.9% of all cases. The results of the linear regression as well as the neural network are worse, since 25% of all cases predict a RR which exceeds the true RR by more than $\theta = 0.2$. In addition, there is no observation where the predicted RR of the mixture regression model exceeds the true RR by more than $\theta = 0.6$. For example, if we estimate a RR of 1, the true value is bigger than 0.4. Thus, for a case where a full recovery is predicted, we know that, at most, 40% of the exposure at default will be recovered. In case of the linear regression model as well as the neural network, there are cases where the predicted RR overestimates the true RR by more than $\theta = 0.9$. We refer to the same example as above. If the estimated RR is 1, the true RR can be smaller than 0.1, which is almost a total loss even though the model predicts a full recovery. Moreover, we notice that the behavior of the linear regression model and the neural network is similar.

The out-of-sample results are similar to the in-sample results. The behavior of the linear regression model equals the behavior of the neural networks. Moreover, regarding the maximum difference, the out-of-sample results for the mixture regression model CMw/oEAD are even slightly better than in-sample, since there is no prediction which overestimates the true RR by the value $\theta = 0.5$. Similar to the in-sample results, the mixture regression model overestimates the true RR by more than $\theta = 0.1$ in 14% of all cases, whereas the linear regression model as well as the neural network exceed the observed RR in 30% of all observations. In addition, the true RR is overestimated by more than $\theta = 0.2$ in only 4.1% of all observations in case of the mixture regression model and in 26% of all observations in case of the linear regression model as well as the neural network. Since the maximum difference is smaller in case of the mixture regression model and the predicted RR exceeds the observed RR by 0.1 resp. 0.2 in only 14% resp. 4.1% of all cases instead of 30% resp. 26% of all cases, we conclude that the mixture regression model CMw/oEAD outperforms the neural network as well as the linear regression model.

## 6. Summary and Conclusions

We compared different models to predict the RR, namely regression methods, decision trees, neural networks, and mixture regression models. Additionally, we investigated how information on an economic crisis can be embedded into the models.

For our analysis, we considered a data set of US-based SMEs obtained from GCD. We use the definition of the workout RR. Empirical RRs exhibit a multimodal structure with three modes at 0, 0.5, and 1. Since earlier studies in the literature point out that an economic crisis during the time to resolution has an impact on the RR, we use a predictive Crisis Indicator (resp. Crisis Probability).

The best models are in-sample as well as out-of-sample the mixture regression models, especially the concomitant variable model which regresses the density as well as the probability that an observation belongs to a certain cluster. We find (by model selection with the BIC) that including the Crisis Probability is preferable compared to including the Crisis Indicator. The neural network outperforms in-sample the linear regression model, but the results are similar out-of-sample. The quantile regression models lead to higher MSEs than the linear regression models. Decision trees performed worst in our study.

Concluding, let us propose some areas for future research in predicting and modeling the RR. In the present study, the RR can take values greater than one as well as smaller than zero and we

conclude that the mixture regression models outperform the other models. Since most of the studies consider observations with RR in the unit interval, the question arises whether the mixture regression models also outperform the other methods in such a restricted data set. Finally, copulas become a popular tool in machine learning and quantile regression (see [39,40], respectively). In particular, a combination of neural networks and regular vine copulas proposed by [41] seems to be promising for modeling recovery rates.

**Author Contributions:** Conceptualization, A.M., M.S., and R.Z.; Data curation, A.S.; Formal analysis, A.S.; Investigation, A.S.; Methodology, A.M., M.S., and R.Z.; Software, A.S.; Supervision, A.M., M.S., and R.Z.; Writing—original draft, A.M., M.S., A.S., and R.Z.; Writing—review & editing, A.M., M.S., A.S., and R.Z. All authors have read and agreed to the published version of the manuscript.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Appendix A. Attributes in This Study

Table A1 gives an overview over all variables from GCD that were used.

**Table A1.** Attributes of the data.

| Attributes | Type | Reference Category | Borrower (B)/Loan (L) Level? | Aggregation for Loan Level |
|---|---|---|---|---|
| Country of Business | categorical | Unknown | B | |
| Public Private Indicator | categorical | Unknown | B | |
| Primary Industry Code | categorical | Unknown | B | |
| Leveraged Finance Indicator | categorical | No | B | |
| Operating Company Indicator | categorical | Unknown | B | |
| Incomplete Portfolio | categorical | No | B | |
| Number of Loans | metric | | B | |
| Nature of Default | categorical | Unknown | B | |
| Lender Issued Amount | categorical | No | B | |
| Entity Sales | metric | | B | |
| EAD | metric | | B | |
| Crisis Indicator resp. Crisis Probability | categorical | No | B | |
| Collateral Indicator | categorical | Unknown | L | at least one collateral |
| Guarantee Indicator | categorical | No | L | at least one guarantee |
| Default Amount | metric | | L | Sum |
| Utilization Rate | metric | | L | Median |
| Guarantor Rating Moody's | categorical | Unknown | L | worst rating |
| Rank of Collateral | categorical | Unknown | L | worst rank |

## References

1. Basel Comittee on Banking Supervision. *Basel II: International Convergence of Capital Measurement and Capital Standards: A Revised Framework*; Bank for International Settlements: Basel, Switzerland, 2004.
2. European Banking Authority. *Guidelines on PD Estimation, LGD Estimation and the Treatment of Defaulted Exposures*; 2017. Available online: https://eba.europa.eu/sites/default/documents/files/documents/10180/2033363/6b062012-45d6-4655-af04-801d26493ed0/Guidelines%20on%20PD%20and%20LGD%20estimation%20(EBA-GL-2017-16).pdf (accessed on 15 September 2020).
3. Ye, H.; Bellotti, A. Modelling Recovery Rates for Non-Performing Loans. *Risks* **2019**, *7*, 19. [CrossRef]
4. Keijsers, B.; Diris, B.; Kole, E. Cyclicality in Losses on Bank Loans. *J. Appl. Econom.* **2018**, *33*, 533–552. [CrossRef]
5. Brumma, N.; Winckle, P. GCD Downturn LGD Study 2017. *SSRN Electron. J.* **2017**. Available online: https://www.globalcreditdata.org/system/files/documents/gcd_downturn_lgd_study_2017.pdf (accessed on 15 September 2020). [CrossRef]

6. Bastos, J. A. *Predicting Bank Loan Recovery Rates with Neural Networks*; CEMAPRE Working Papers; Centre for Applied Mathematics and Economics (CEMAPRE), School of Economics and Management (ISEG), Technical University of Lisbon: Lisbon, Portugal, 2010.

7. Yao, X.; Crook, J.; Andreeva, G. Support Vector Regression for Loss Given Default Modelling. *Eur. J. Oper. Res.* **2015**, *240*, 528–538. [CrossRef]

8. Dermine, J.; Carvalho, D.C.N. Bank Loan Losses-Given-Default: A Case Study. *J. Bank. Financ.* **2006**, *30*, 1219–1243. [CrossRef]

9. Qi, M.; Zhao, X. Comparison of Modeling Methods for Loss Given Default. *J. Bank. Financ.* **2011**, *35*, 2842–2855. [CrossRef]

10. Bellotti, A.; Crook, J. Loss Given Default Models for UK Retail Credit Cards. *Building* **2009**, 1–28. Available online: https://www.researchgate.net/publication/215991287_Loss_Given_Default_models_for_UK_retail_credit_cards (accessed on 15 September 2020).

11. Gupton, G.; Stein, R. Losscalc v2: Dynamic Prediction of LGD-Modeling Methodology. Moody's/KMV 2005. Available online: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.87.7197 (accessed on 15 September 2020).

12. Loterman, G.; Brown, I.; Martens, D.; Mues, C.; Baesens, B. Benchmarking Regression Algorithms for Loss Given Default Modeling. *Int. J. Forecast.* **2012**, *28*, 161–170. [CrossRef]

13. Krüger, S.; Rösch, D. Downturn LGD Modeling using Quantile Regression. *J. Bank. Financ.* **2017**, *79*, 42–56. [CrossRef]

14. Gostkowski, M; Gajowniczek, K. Weighted Quantile Regression Forests for Bimodal Distribution Modeling: A Loss Given Default Case. *Entropy* **2020**, *22*, 545. [CrossRef]

15. Tomarchio, S.D.; Punzo, A. Modelling the Loss Given Default Distribution via a Family of Zero-and-one Inflated Mixture Models. *J. R. Stat. Soc. Ser. A* **2019**, *182*, 1247–1266. [CrossRef]

16. Altman, I.E.; Kalotay, A.E. Ultimate Recovery Mxtures. *J. Bank. Financ.* **2014**, *40*, 116–129. [CrossRef]

17. Wang, H.; Forbes, C.S.; Fenech, J.-P.; Vaz, J. *The Determinants of Bank Loan Recovery Rates in Good Times and Bad–New Evidence*; Monash Econometrics and Business Statistics Working Papers; 2018. Available online: https://ideas.repec.org/p/msh/ebswps/2018-7.html (accessed on 15 September 2020).

18. Calabrese, R. Predicting Bank Loan Recovery Rates with a Mixed Continuous-Discrete Model. *Appl. Stoch. Model. Bus. Ind.* **2012**, *30*, 99–114. [CrossRef]

19. Calabrese, R. Downturn Loss Given Default: Mixture Distribution Estimation. *Eur. J. Oper. Res.* **2014**, *237*, 271–277. [CrossRef]

20. Fahrmeir, L.; Kneib, T.; Lang, S.; Marx, B. *Regression: Models, Methods and Applications*; Springer: Berlin, Germnay, 2013.

21. Hosmer, D.W.; Lemeshow, S. *Applied Logistic Regression*; John Wiley and Sons: Hoboken, NJ, USA, 2013.

22. Günther, F.; Fritsch, S. Neuralnet: Training of Neural Networks. *R J.* **2010**, *2*, 30–38. [CrossRef]

23. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer: New York, NY, USA, 2001.

24. Sarle, W.S. Neural Networks and Statistical Models. In Proceedings of the Nineteenth Annual SAS Users Group International Conference, Dallas, TX, USA, 10–13 April 1994.

25. Riedmiller, M.; Braun, H. A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm. *IEEE Int. Conf. Neural Netw.* **1993**, *1*, 586–591.

26. Frühwirth-Schnatter, S. *Finite Mixture and Markov Switching Models*; Springer: New York, NY, USA, 2006.

27. Grün, B.; Leisch; F. Fitting Finite Mixtures of Linear Regression Models with Varying and Fixed Effects in R. In *Proceedings in Computational Statistics*; Rizzi, A., Vichi, M., Eds.; Physica Verlag: Heidelberg, Germnay, 2006; pp. 853–860.

28. Leisch, F. Flexmix: A General Framework for Finite Mixture Models and Latent Class Regression in R. *J. Stat. Softw. Artic.* **2004**, *11*, 1–18. [CrossRef]

29. Murphy, K.P. *Machine Learning: A Probabilistic Perspective*; MIT Press: Cambridge, MA, USA, 2012.

30. Grün, B.; Kepler, J.; Leisch, F. Flexmix Version 2: Finite Mixtures with Concomitant Variables and Varying and Constant Parameters. *J. Stat. Softw.* **2008**, *28*, 1–35. [CrossRef]

31. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. R. Stat. Society. Ser. B (Methodol.)* **1977**, *39*, 1–38.

32. Höcht, S.; Zagst, R. *Loan Recovery Determinants: A Pan-European Study*; Working Paper; 2008. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2443724 (accessed on 15 September 2020).

33. Bastos, J.A. Forecasting Bank Loans Loss-Given-Default. *J. Bank. Financ.* **2010**, *34*, 2510–2517. [CrossRef]

34. Ernst, C.; Grossmann, M.; Hocht, S.; Minden, S.; Scherer, M.; Zagst, R. Portfolio Selection under Changing Market Conditions. *Int. J. Financ. Serv. Manag.* **2009**, *4*, 48–63. [CrossRef]

35. Panagiotopoulou, K. Modeling and Forecasting Downturn LGD. Master's Thesis, Technische Universität München, Munich, Germany, 2018.

36. Schwarz, G.E. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464. [CrossRef]

37. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **1974**, *19*, 716–723. [CrossRef]

38. Lantz, B. *Machine Learning with R*, 2nd ed.; Packt Publishing: Birmingham, UK, 2015.

39. Kraus, D.; Czado, C. D-vine Copula Based Quantile Regression. *Comput. Stat. Data Anal.* **2017**, *110*, 1–18. [CrossRef]

40. Elidan, G. Copulas in Machine Learning. In *Copulae in Mathematical and Quantitative Finance*; Jaworski, P., Durantem, F., Härdle, W., Eds.; Physica Verlag: Heidelberg, Germnay, 2013; pp. 39–60.

41. Zhang, S.; Geng, B.; Varshney, P.; Rangaswamy, M. Fusion of Deep Neural Networks for Activity Recognition: A Regular Vine Copula Based Approach. In Proceedings of the 2019 22th International Conference on Information Fusion (FUSION), Ottawa, ON, Canada, 2–5 July 2019.