Peer-reviewed version available at Risks 2019, 7, 19; doi:10.3390/risks7010019

Article

Modelling Recovery Rates for Non-performing Loans

Hui Ye [1,‡] and Anthony Bellotti [1,†,‡*]

[1] Department of Mathematics, Imperial College London
* Correspondence: a.bellotti@imperial.ac.uk
† Current address: Department of Mathematics, Imperial College London, South Kensington, London SW7 2AZ, UK
‡ These authors contributed equally to this work.

Abstract: Based on a rich data set of recoveries donated by a debt collection business, recovery rates for non-performing loans taken from a single European country are modelled using linear regression, linear regression with Lasso, beta regression and inflated beta regression. We also propose a two-stage model: beta mixture model combined with a logistic regression model. The proposed model allows us to model the multimodal distribution we find for these recovery rates. All models are built using loan characteristics, default data and collections data prior to purchase by the debt collection business. The intended use of the models is to estimate future recovery rates for improved risk assessment, capital requirement calculations and bad debt management. They are compared using a range of quantitative performance measures under K-fold cross validation. Among all the models, we find that the proposed two-stage beta mixture model performs best.

Keywords: recovery rates; beta regression; credit risk

1. Introduction

In Basel II, an internal ratings-based approach (the IRB approach) was proposed by the Basel Committee in 2001 to determine capital requirements for credit risk [1]. This IRB approach grants banks permission to use their own risk models or assessments to calculate regulatory capital. Under the IRB approach, banks are required to estimate the following risk components: probability of default (PD), loss given default (LGD), exposure at default (EAD) and maturity (M) [1]. Since Basel II's capital requirement calculation depends heavily on LGD, financial institutions have put more emphasis on modelling LGD in recent years. Unlike the estimation of PD, which is well-established, LGD is not so well-understood and still subject to research. Improving LGD modelling can help financial institutions assess their risk and regulatory capital requirement more precisely, as well as improving debt management.

LGD is defined as the proportion of money financial institutions fail to collect during the collection period, given the borrower has already defaulted. Conversely, Recovery Rate (RR) is defined as the proportion of money financial institutions successfully collected minus the administration fees during the collection period, given the borrower has already defaulted. Equations 1 and 2 give formal definitions of RR and LGD respectively:

- Suppose individual i has already defaulted on a loan, let EAD_i be the Exposure at Default for this individual i.
- Let A_i be the administration costs (e.g., letters, phone calls, visits, lawyers and legal work) incurred for individual i.
- Let R_i be the amount recovered for individual i.

Then,

$$\text{Recovery Rate} = \frac{R_i - A_i}{EAD_i} = \frac{\sum \text{Collections} - \sum \text{Admin Fee}}{\text{Outstanding Balance at Default}} \tag{1}$$

and,

$$\text{Loss Given Default} = 1 - \text{Recovery Rate} = 1 - \frac{R_i - A_i}{EAD_i} \tag{2}$$

RR mainly lies in the interval [0, 1] and typically has high concentrations at the boundary points 0 and 1. It is possible for RR to be negative if recoveries are less than administration costs, $A_i > R_i$, and greater than 1 if recoveries exceed exposure plus administration costs, $R_i > EAD_i + A_i$. Typically, however, RR is truncated within the interval [0,1] when developing LGD models.

The main challenge in estimating LGD, in general, is the bimodal property with high concentrations at 0 and 1 typically present in LGD empirical distributions, where people either repay in full or repay nothing. For the data set we use in this study, we find our LGD distribution is actually tri-modal. Therefore, regression models have been studied that specifically deal with this problem. For example, [2] built Tobit and decision tree models along with beta and fractional logit transformation of the RR response variable to forecast the LGD based on a data set of 55,000 defaulted credit cards in the UK from 1999 to 2005. They conclude that ordinary least squares regression with macroeconomic variables performed the best in terms of forecast performance. [3] proposed a mixed continuous-discrete model, where the boundary values 0 and 1 are modelled by Bernoulli random variables and the continuous part of the RR is modelled by a Beta random variable. This model is then applied to predict RR of Bank of Italy's loans from 1985 to 1999. The result is compared with Papke and Wooldrige's fractional response model with log-log, logistic and complementary log-log link functions [4] and linear regression. The mixed continuous and discrete model achieves the best performance . [5] applied four linear models: ordinary least squares regression, fractional response regression, inverse Gaussian regression, and inverse Gaussian regression with beta transformation and two non-linear models: regression tree and neural network to model the LGD of 3751 defaulted bank loans and bonds in the US from 1985 to 2008. They conclude that fractional response regression is slightly better than the ordinary least squares regression. Moreover, they report that non-linear models perform best. [6] perform a benchmark study of LGD by comparing twenty-four different models using six data sets extracted from international banks. They conclude that non-linear models, such as neural network, support vector machine and mixture models perform better than linear models .

For this project, we specifically model and predict RR for data from a single European country provided by a debt collection company. Due to reasons of commercial confidentiality and data protection, the debt collection company will remain anonymous and some aspects of the data will also be anonymized, including the country of origin. Consequently the data cannot be made publicly available. We apply some of the models that have already been studied previously and also extend the existing models, proposing a new beta mixture model to improve the accuracy of RR prediction. A good prediction of RR will help the debt collection company to determine collection policy for new debt portfolios. It is important to note that the RR we model is different from most RR, as the data only contains positive repayments and no administration fee is recorded. Therefore, all the RRs in our data lie in the range (0, 1] instead of [0, 1]. Figure 1 shows a histogram of RR for the data. We can clearly see that there are modes at 0, 0.55 (approximately) and a high spike at boundary value 1. Since the shape of the empirical RR distribution demonstrates a trimodal feature, it is reasonable to assume that the recovery rate is a mixed type random variable. The multi-modality of RR is a natural consequence of different groups of bad debts being serviced using different strategies; eg one strategy may be that some bad debts are allowed to be written off if the debtor paid back some agreed fixed percentage of the outstanding balance. Having outcome RR within $(0, 1]$ motivates the use of the beta regression model and the multi-modal nature of RR motivates the use of a mixture model within this context.

The beta mixture model has been applied successfully within several other application domains. [7] show how to apply the beta mixture regression model in several bioinformatics applications such as meta-analysis of gene expression data and to cluster correlation coefficients between gene expressions. [8] use a beta mixture model to describe DNA methylation patterns, helping to reduce the dimensionality of microarray data. [9] use a beta mixture model as the basis of an anomaly detection

system. Their network data is typically bounded which suggests a beta distribution, and the use of the beta mixture allows them to identify latent clusters in normal network use.

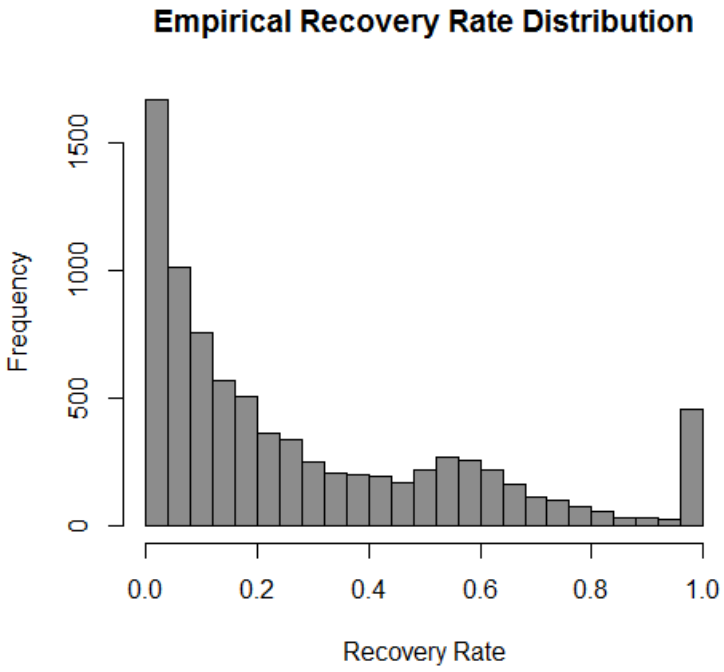


Figure 1. Histogram of Recovery Rates for 8237 loans after pre-preprocessing described in Section 2. The stack of 1s shows frequency of $RR = 1$, but the stack at 0 shows frequency for small $RR > 0$.

Inspired by Calabrese's mixed continuous-discrete model [3], we propose a two-stage model composed of:

- A beta mixture model parameterized by mean and precision based on two sets of predictor variables on the interval of (0, 1) in order to model the two modes located at just after 0 and around 0.55.
- A logistic regression model for the mode at boundary value 1.

The above proposed model allows representation of the trimodal feature of the data. The beta mixture component groups the clients into two clusters for RR<1, based on their personal information, debt conditions and repayment history, which may become useful information for other business analysis and decision-making, then use logistic regression to model the third case RR=1. In addition, we will also use linear regression, linear regression with Lasso, beta regression and inflated beta regression to model RR. Model performance will be measured by mean squared error, mean absolute error and mean aggregate absolute error under K-fold cross validation.

To our knowledge, this is the first study for estimating RR for portfolios of non-performing loans using a statistical model, and the first use of a beta mixture model for LGD. We also develop a novel procedure for predicting an expected value of outcome from a beta mixture model based on assigning a new observation to one of the clusters in the mixture. The remainder of the article is organized as follows: Section 2 provides a detailed data overview. Section 3 introduces the modelling methodology with great emphasis on the proposed beta mixture model combined with logistic regression model. Section 4 analyzes some important features of the models and reports the model performance and the last section concludes with key findings and future recommendations.

2. Data

Three data sets are provided by the debt collection company:

Data set 1 provides 48 predictor variables of personal information including socio-demographic variables, Credit Bureau Score and debt status for 120699 individuals for loans originating between January 1998 and May 2014 from several different financial institutions. 97.5% of them have credit card debt and only 2.5% have mortgage debt. Partial information was extracted from a Bad Debt Bureau. Each record corresponds to a bad loan and has a unique key Loan.Ref.

Data set 2 records all the recoveries made by the bank *before* the debt collection company purchased the debt portfolio. It contains 15 predictor variables about historical collection information, which includes number of calls, contacts and visits made by the bank to collect the debt. It also includes repayments in the format of monthly summary. In total, there are 42832 individuals' records in data set 2, among which only 34807 individuals can be matched to data set 1 by Loan.Ref. Numbers of calls, contacts, visits, repayment and some other monthly activities are aggregated by summing for each loan identified by Loan.Ref.

Data set 3 records all the recoveries made by the debt collection company *after* they purchased the debt portfolio from the bank. It includes 12 predictor variables about the ongoing collection information. There are 8281 individuals in total, among which only 8237 individuals are from data set 1. Since only positive repayments are recorded, all the recovery rates we calculated are strictly greater than 0. Therefore, in the modelling section, we will only focus on the recovery modelling in the interval $(0, 1]$, which is slightly different from the usual RR defined in $[0, 1]$. The debt collection period recorded in this data set is from January 2015 to end of November 2016.

Figure 2 shows how the data has been joined. There are 8237 data points presented in data set 3, but only 7161 individual historical collection information are recorded in data set 2. In these cases, there are no historical recoveries by bank, i.e., no calls, contacts, visits or payments for the remaining 1076 individuals. Therefore, a value of 0 will be assigned to aggregate recoveries in data set 2 for the remaining 1076 individuals. The modified data set 2 is then joined to data set 1 and 3 by the unique key *Loan.Ref* and we obtain a table of 8237 data points with 61 variables.

Table A1 gives descriptive statistics for each of the variables in the joined data set used in the statistical modelling. The predictor variable Pre-Recovery Rate is the bank's RR before the debt portfolio was purchased. The minimum value is -0.130, which is negative due to the substantial amount of administration fee exceeding repayments incurred during the collection period. The predictor variable Credit Bureau Score is a generic credit score provided by a credit bureau.

2.1. Recovery Rate Calculation

Since the repayments in data sets 2 and 3 are recorded in the format of monthly activity summaries, each individual may have several repayments for the same loan. Therefore, we define the recovery rate as the sum of repayments minus the administration fee (if available) over the original balance of the loan, which is also equivalent to the difference between original balance and ending balance over the original balance. For each individual i, RR is calculated using:

$$\text{Recovery Rate}_i = \frac{\sum \text{Repayments}_i - \text{AdminFee}_i}{\text{Original Balance}_i} = \frac{\text{Original Balance}_i - \text{Ending Balance}_i}{\text{Original Balance}_i} \quad (3)$$

Figure 1 is the empirical RR histogram calculated based on Equation 3, for the 8237 data points after pre-processing. The remainder of 112462 data points not included in the anslysis essentially have RR=0, but we do not know whether they have been serviced or not, so they are not included in the analysis. Essentially, the goal of our model is to estimate RR computed from data set 3 (post-purchase), based on pre-purchase information given in data sets 1 and 2.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 February 2019 doi:10.20944/preprints201902.0135.v1

Peer-reviewed version available at Risks 2019, 7, 19; doi:10.3390/risks7010019

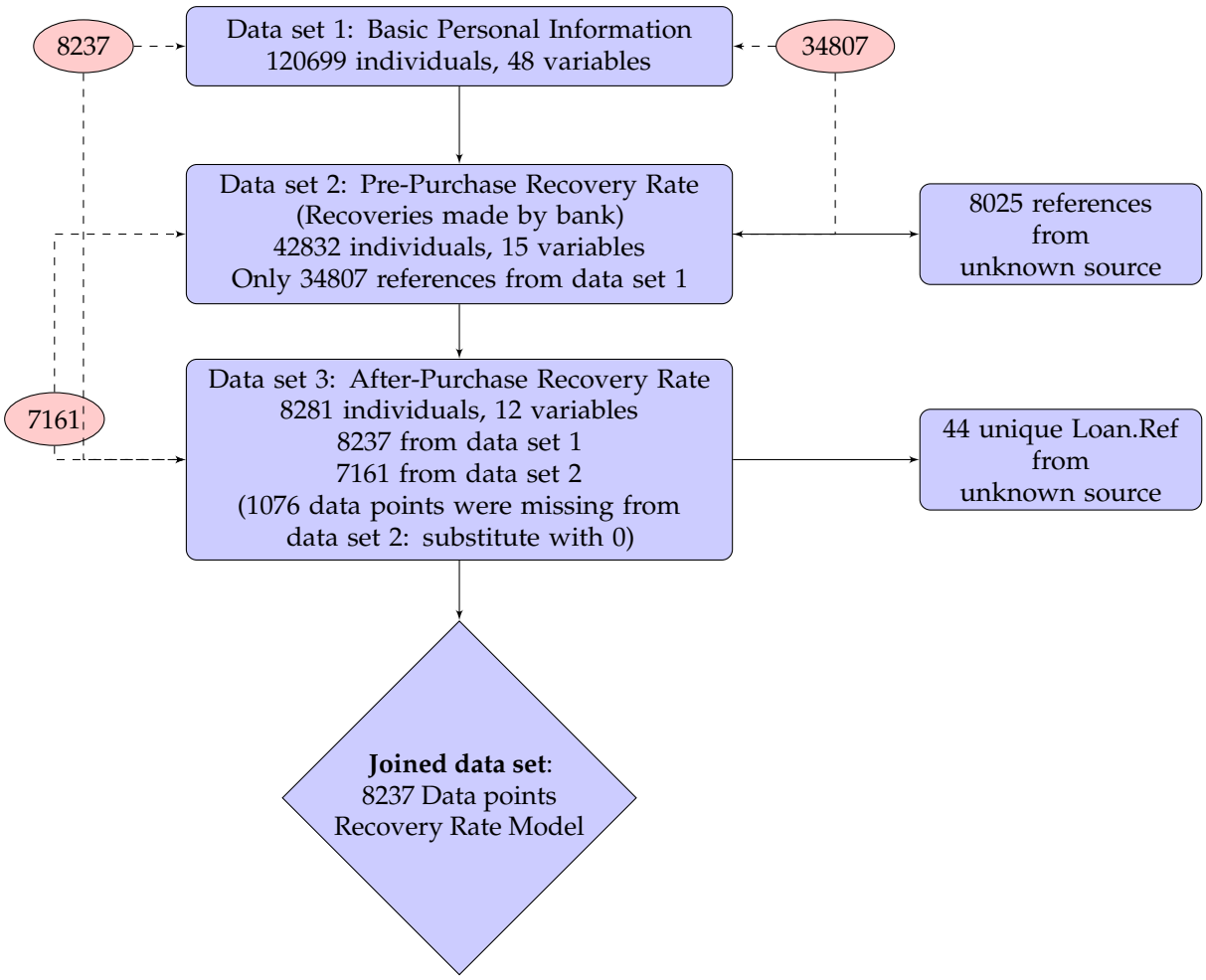


Figure 2. Joining the three data sets

3. Modelling Methodology

We apply various models to estimate RR. In all cases, model performance will be measured within a K-fold cross validation framework. We first try using ordinary least squares linear regression, with and without stepwise backward variable selection using the AIC criterion. In the following sub-sections, we list the other modelling approaches we explore. Let y indicate the outcome variable, recovery rate, and X is a corresponding vector of predictor variables.

3.1. Linear regression with Lasso

We apply linear regression with a Lasso (Least Absolute Shrinkage and Selection Operator) penalty. The model structure is

$$y = \beta_0 + \boldsymbol{\beta}^T X + \epsilon$$

where β_0 and $\boldsymbol{\beta}$ are intercept and coefficients to be estimated and ϵ is the error term. Then, estimation using least squares error with Lasso is given by the optimization problem on a training data set of N observations:

$$(\hat{\beta_0}, \hat{\boldsymbol{\beta}}) = \underset{\beta_0, \boldsymbol{\beta}}{\operatorname{argmin}} \left[\frac{1}{N} \sum_{i=1}^{N} \left(y_i - \beta_0 - \boldsymbol{\beta}^T X_i \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right], \tag{4}$$

where $\lambda > 0$ is a tuning parameter controlling the size of regularization. Regression with Lasso will tend to shrink coefficient estimates to zero and hence is a form of variable selection [10]. The value of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 February 2019 doi:10.20944/preprints201902.0135.v1

Peer-reviewed version available at Risks 2019, 7, 19; doi:10.3390/risks7010019

λ is chosen using K-fold cross validation. For this project, the R packages 'lars' [11] and 'glmnet' [10] are used to estimate linear regression with Lasso.

3.2. Multivariate Beta Regression

The problem with linear regression is that it does not take account of the particular distribution of RR which is between 0 and 1. The beta distribution, with two shape parameters α and β, allows us to model RR in the open interval $(0, 1)$:

$$f(y_i; \alpha_i, \beta_i) = \frac{\Gamma(\alpha_i + \beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)} y_i^{\alpha_i - 1}(1 - y_i)^{\beta_i - 1}, \quad 0 < y_i < 1, \tag{5}$$

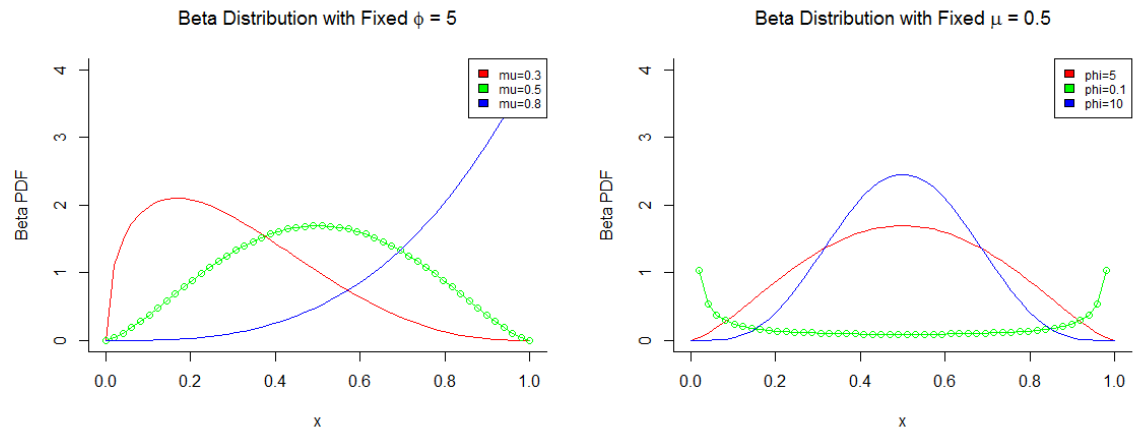
where α, $\beta > 0$ are the shape parameters and $\Gamma(\cdot)$ is the Gamma function. The beta distribution is reparametrized by mean and precision parameters, denoting by μ and ϕ respectively, following [12], since this parameterization meaningfully express the expected value and variance:

$$\phi_i = \alpha_i + \beta_i, \quad E(y_i) = \mu_i = \frac{\alpha_i}{\alpha_i + \beta_i}, \quad Var(y_i) = \frac{\mu_i(1 - \mu_i)}{\phi_i + 1}, \tag{6}$$

The reparameterized beta distribution is then

$$f(y_i; \mu_i, \phi_i) = \frac{\Gamma(\phi_i)}{\Gamma(\mu_i\phi_i)\Gamma((1 - \mu_i)\phi_i} y_i^{\mu_i\phi_i - 1}(1 - y_i)^{(1-\mu_i)\phi_i - 1}, \quad 0 < y_i < 1, \tag{7}$$

with $0 < \mu_i < 1$ and $\phi_i > 0$. Figure 3a demonstrates three examples of the beta distribution with fixed $\phi = 5$ and different μ. The variance is maximized at $\mu = 0.5$. Figure 3b demonstrates another three examples of beta distribution with fixed $\mu = 0.5$ and different ϕ. The precision parameter ϕ



(a) Beta Distribution with Fixed ϕ

(b) Beta Distribution with Fixed μ

Figure 3. Beta Distribution

is negatively correlated with $Var(y_i)$, given a fixed μ. Furthermore, the variance of Y is a function of μ, which enables the regression to model heteroskedasticity. RR is modelled as $y_i \sim B(\mu_i, \phi_i)$ for $i \in (1, \cdots, N)$ for sample size N. The multivariate beta regression model [13] is defined as:

$$F_1(\mu_i) = \eta^T X_i = \xi_{1i},$$

$$F_2(\phi_i) = \gamma^T W_i = \xi_{2i},$$

where η is a vector of parameters which needs to be estimated corresponding to predictor variables X and γ is a vector of parameters which needs to be estimated corresponding to predictor variables W. The predictor variables in W may be the same as in X, or a subset, or contain different variables. For this study, W will have a subset of predictor variables determined using stepwise variable selection. The link function ensures that $\mu_i \in (0,1)$ and $\phi_i > 0$. We apply Logit and Log link function to μ_i and ϕ_i respectively:

$$\mu_i = \frac{1}{1+e^{-\eta^T X_i}}, \quad \phi_i = e^{-\gamma^T W_i}.$$

With this multivariate beta regression model, η and γ can be estimated by maximum likelihood estimation, where the log-likelihood function is

$$\begin{aligned} L(\eta,\gamma) = \sum_{i=1}^{N} \Big[& \log\Gamma(\phi_i) - \log\Gamma(\mu_i\phi_i) - \log\Gamma((1-\mu_i)\phi_i) \\ & + (\mu_i\phi_i - 1)\log y_i + ((1-\mu_i)\phi_i - 1)\log(1-y_i)\Big]. \end{aligned} \tag{8}$$

By substituting $\mu_i = F_1^{-1}(\eta^T X_i)$ and $\phi_i = F_2^{-1}(\gamma^T W_i)$ into equation (8), the log-likelihood is obtained as a function of η and γ. The parameters can be estimated using Broyden-Fletcher-Goldfarb-Shanno (BFGS) quasi-Newton method, which is considered to be the most appropriate method [14,15].

3.3. Inflated Beta Regression

The disadvantage of beta regression is that it does not include the boundary values 0 or 1. Therefore, a modification is required before fitting the model. In order to better represent RR on the boundaries 0 and 1, [3] suggested to consider RR as a mixture of Bernoulli random variables for the boundary 0 and 1, and a Beta random variable for the open interval (0,1). The distribution for this inflated beta regression on [0,1] is then defined as

$$f_Y(y) = \begin{cases} p_0, & \text{if } y = 0 \\ (1-p_0-p_1) f_B(y;\alpha,\beta), & \text{if } 0 < y < 1 \\ p_1, & \text{if } y = 1 \end{cases} \tag{9}$$

for y $\in$ [0,1], $p_0 = P(y=0)$, $p_1 = P(y=1)$, $0 < p_0 + p_1 < 1$ and $f_B(y)$ is the beta distribution defined in Section 3.2. Moreover, if RR $y \in (0,1]$, ie it only inflates at one, as our data does, then the distribution is just

$$f_Y(y) = \begin{cases} (1-p_1) f_B(y;\alpha,\beta), & \text{if } 0 < y < 1 \\ p_1, & \text{if } y = 1 \end{cases} \tag{10}$$

We use maximum likelihood estimation to estimate parameters for Bernoulli random variable and Beta random variables, parametrizing the discrete part in the following way [3]:

$$s_i = \frac{p_1}{p_1+p_0}, \quad d_i = p_0 + p_1,$$

The log-likelihood function is then

$$\begin{aligned} L(s,d,\alpha,\beta) = & \sum_{y_i=0} \log(1-s_i) + \sum_{y_i=0} \log(d_i) + \sum_{y_i=1} \log(s_i) + \sum_{y_i=1} \log(d_i) \\ & + \sum_{0<y_i<1} \log(1-d_i) + \sum_{0<y_i<1} \log(f_B(y;\alpha_i,\beta_i)). \end{aligned} \tag{11}$$

The continuous beta random variables can be parametrized in the same way as described in Section 3.2.

Peer-reviewed version available at Risks 2019, 7, 19; doi:10.3390/risks7010019

3.4. Beta Mixture Model combined with Logistic Regression

Examining the distribution of RR shown in Figure 1, it can be seen that the distribution between 0 and 1 is bimodal. For this reason, we consider a beta mixture model to deal with what appears to be two different groups of recoveries. We propose a two-stage model: beta mixture model combined with logistic regression. The beta mixture model allows us to model the multimodality of RR in the interval (0,1). This is similar to the two-stage (decision tree) model used by [2], but with a beta mixture used for regression.

Firstly, classify RR into ones and non-ones using logistic regression. Secondly, within the non-ones group, a mixture of beta distributions is used to model RR in the range (0,1). In general, a mixture of beta distribution consists of m components where each component follows a parametric beta distribution. The prior probability of component j is denoted as π_j, where $j \in (1, \cdots, m)$. Let M_j denote the jth component/cluster in the beta mixture model. The beta mixture model with m components is defined as:

$$\begin{aligned} g(y;\mu,\phi) &= \sum_{j=1}^{m} \pi_j f_j(y; X, \mu_j, \phi_j) \\ &= \sum_{j=1}^{m} \pi_j f_j(y; X, W, F_1^{-1}(\eta_j^T X_i), F_2^{-1}(\gamma_j^T W_i)) \\ &= \sum_{j=1}^{m} \pi_j f_j(y; X, W, \eta_j, \gamma_j), \end{aligned}$$

where f_j is the beta distribution corresponding to the jth component with separate parameter vectors η_j and γ_j. The same link functions are used as in Section 3.2. The prior probabilities, π_j need to satisfy the following conditions:

$$\sum_{j=1}^{m} \pi_j = 1, \quad \pi_j \geq 0.$$

The iterative Expectation–Maximization (EM) algorithm is used to estimate the parameters of the beta mixture model as described by [16]. In particular, R package 'flexmix' [16–18] embedded in R package 'betareg' [13,19] is applied to estimate the model. Figure 4 illustrates the two-stage mixture model as a decision tree.

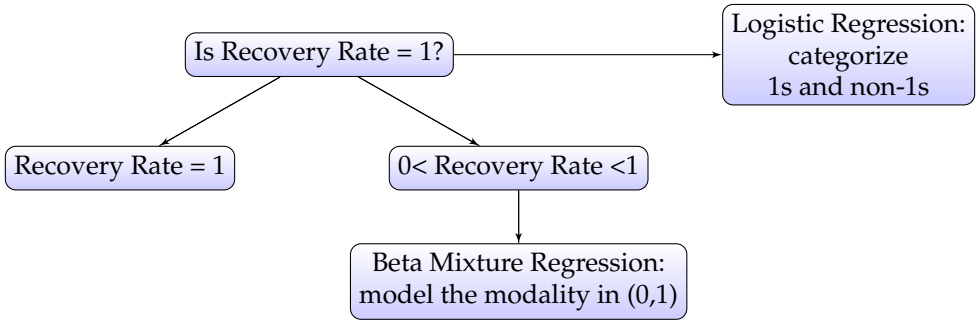


Figure 4. Estimate the expected value of RR using two-stage decision tree model.

The choice of m in the model depends on the number of clusters expected in the data. From our analysis of the recoveries for the data set we are using, $m = 2$ will be used since this corresponds to the two modes we see in the RR distribution for RR<1 as shown in Figure 1. If it was not clear how many clusters may exist, approaches can be used based on AIC...

Peer-reviewed version available at Risks 2019, 7, 19; doi:10.3390/risks7010019

3.4.1. Predictions using the Beta Mixture Model

Given the beta mixture model, we need to predict the RR for new clients based on their information, i.e., X_{new} and W_{new}. Figure 5 shows a flowchart explaining how to calculate the estimated RR from the beta mixture model. This gives an expected value of RR y conditional on the cluster M_j. Therefore, we need to first identify which cluster the new observation belongs to. Even though the R package 'betareg' [13,19] can compute the conditional expectation for us, it does not identify which cluster the new points should be assigned to. Therefore, we propose a method to do this. In general, there are two feasible approaches to assign a new observation to M_j:

1. Assign the new observation to the cluster that achieves the highest log-likelihood. This is a hard clustering approach which assigns the observation to exactly one cluster [20].
2. Assign the new observation to each cluster j with probability $P(M_j)$. This is a soft clustering approach which assigns the observation to a percentage weighted cluster [16].

Decomposing the expected value of y using the Law of Total Expectation, we get

$$E(y \mid x_i) = \sum_{j=1}^{m} P(M_j|x_i)E(y|x_i, M_j) \tag{12}$$

where $E(y|x_i, M_j)$ is calculated from the beta mixture model prediction (refer to Figure 5). We can replace $P(M_j|x_i) = \frac{f(x_i|M_j)P(M_j)}{f(x_i)}$ where $f(x_i) = \sum_{j=1}^{m} f(x_i|M_j)P(M_j)$, to get

$$E(y \mid x_i) = \frac{\sum_{j=1}^{m} f(x_i|M_j)P(M_j)E(y|x_i, M_j)}{f(x_i)} \tag{13}$$

where $P(M_j)$ is the prior probability of belonging to cluster M_j. The density $f(x_i|M_j)$ is estimated using kernel density estimation,

$$\hat{f}(x^{\text{new}}) = \sum_{i=1}^{n} \frac{1}{n\prod_{k=1}^{d} h_{i,k}} \prod_{k=1}^{d} K(\frac{x_k^{\text{new}} - x_{i,k}}{h_{i,k}})$$

where $K(\cdot)$ is the Gaussian kernel [21] and d is the number of dimensions in data x. In addition, x_i may be high-dimensional, which makes the kernel density estimation computationally expensive. As a remedy, we applied Principal Component Analysis (PCA) to reduce the dimension of x_i, then kernel density estimation is performed in the reduced dimension space.

Approach 1: Maximum log-likelihood. Given a new observation x_i, choose j that maximizes the density:

$$\arg\max_j \quad \log f(y|x_i, M_j),$$

which is computed using the log-likelihood function. If the objective function is maximized with respect to Cluster M_j, then set

$$P(M_j|x_i) = 1 \text{ and } P(M_k|x_i) = 0 \text{ for all } k \neq j$$

and hence, from Equation 12, the expected value of y is given by $E(y|x_i, M_1)$.

Approach 2: Prior Probability. Treat $P(M_j)$ as a prior estimated using methods given in Table 1 and use in Equation 13 for soft clustering. By substituting $P(M_j)$ given in Table 1 into Equation 13, we can compute $E(y \mid x)$ for $y \in (0,1)$.

After calculating $E(y \mid x)$ for the interval (0,1) using the beta mixture model, the boundary 1 needs to be taken into consideration using a logistic regression model. From the decision tree defined

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 February 2019 doi:10.20944/preprints201902.0135.v1

Peer-reviewed version available at Risks 2019, 7, 19; doi:10.3390/risks7010019

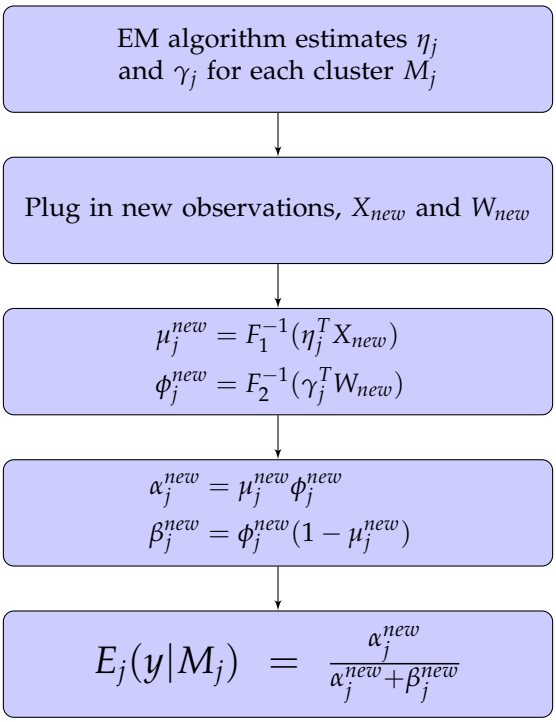


Figure 5. Prediction of RR conditional on each cluster M_j

Table 1. Determining $P(M_j)$ in Approach 2

Approach 2	$P(M_1)$	$P(M_2)$
π_j prior	Extract π_j from the EM algorithm $\pi_j[1]$	Extract π_j from the EM algorithm $\pi_j[2]$
Prior based on training set cluster size ratio	$\frac{\text{Cluster 1 size}}{\text{Total sample size}}$	$\frac{\text{Cluster 2 size}}{\text{Total sample size}}$
Indifferent Prior	$\frac{1}{2}$	$\frac{1}{2}$

in Figure 4, the logistic regression can provide the estimates at the first leaf node: $P(y = 1 \mid X = x)$. Then, the overall expectation of RR $y \in (0,1]$ is

$$\begin{aligned} E(y \mid x) &= P(y=1 \mid x)E(y \mid x, y=1) \quad + \quad P(0<y<1 \mid x)E(y \mid x, 0<y<1) \\ &= P(y=1 \mid x) \times 1 \quad + \quad (1-P(y=1 \mid x))\, E(y \mid x, 0<y<1) \\ &= P(y=1 \mid x) \quad + \quad (1-P(y=1 \mid x))\, E(y \mid x, 0<y<1) \end{aligned} \tag{14}$$

where $E(y \mid x, 0 < y < 1)$ is the predicted RR from the beta mixture model using Approach 1 or 2.

4. Results

The linear model has an adjusted R^2 of 0.69, which is considerably higher than most models of RR (eg see [2,6]), and this can be explained by the richness of data, especially collections information. We expect that the linear regression model is misspecified, due to the range of the outcome variable and this is confirmed in the residual vs. fitted plot for the model and a Breusch-Pagan test for heteroscedasticity ($p < 0.0001$).

For the beta mixture model, we use all variables for X, but variable selection for W based firstly on the output of stepwise selection using AIC in linear regression and then on a series of likelihood ratio tests. The result is the selection of four variables for W: pre-recovery rate, post balance, customer

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 February 2019 doi:10.20944/preprints201902.0135.v1

Peer-reviewed version available at Risks 2019, 7, 19; doi:10.3390/risks7010019

payment frequency and credit bureau score. Table 2 shows parameter estimates for η and γ for the two clusters, along with coefficient estimates under standard beta regression in the interval (0, 1) for comparison. In Table 2, there are 'NA' values for some of the p-values in the beta mixture model. This

Table 2. η, γ estimated by EM algorithm.
M1 and M2 represent Clusters 1 and 2

	Beta Mixture Model in (0,1)				Beta Regression in (0,1)	
Variables	**M1 Estimate**	**Pr(>|z|)**	**M2 Estimate**	**Pr(>|z|)**	**Betareg Estimate**	**Pr(>|z|)**
η						
(Intercept)	-0.67015	<0.0001	-2.62862	<0.0001	-1.80064	<0.0001
Product R	-0.03376	0.47711	-0.00766	0.59733	0.02270	0.41766
Principal	0.00056	NA	0.00114	NA	0.00081	0.00000
Interest	0.00065	<0.0001	0.00118	NA	0.00097	0.00000
Insurance	0.00082	<0.0001	0.00116	<0.0001	0.00086	<0.0001
Late Charges	0.00042	0.00578	0.00115	<0.0001	0.00072	<0.0001
Overlimit Fees	-0.00105	0.07594	0.00145	<0.0001	0.00018	0.52533
Credit limit	0.00004	NA	-0.00001	NA	-0.00003	<0.0001
Sex = Male	0.03659	0.17453	-0.01412	0.13364	0.00969	0.43796
Marital status =						
Divorced	-0.01175	0.85305	-0.01427	0.47359	-0.03144	0.25840
Married	-0.06356	0.10819	-0.01476	0.16836	-0.03850	0.01957
Single	0.00982	0.83178	0.00695	0.63324	0.00332	0.86926
Widow	-0.14627	0.19497	0.02311	0.51404	-0.03869	0.45314
Other	0.12328	0.17954	-0.03476	0.22384	0.04570	0.24125
Age	-0.00273	0.05378	-0.00038	0.42389	-0.00115	0.07159
Credit Bureau Score	0.00059	0.10337	0.00007	0.07890	0.00038	0.00222
Bureau bad debt	-0.32990	0.01290	-0.06936	<0.0001	-0.24123	0.00000
Cust Payment Freq	0.06530	<0.0001	0.03506	<0.0001	0.05046	<0.0001
Post Balance	-0.00106	NA	-0.00127	NA	-0.00103	0.00000
Total Paid Amount	0.00004	NA	-0.00038	NA	-0.00014	<0.0001
Total Calls	-0.00044	0.00515	-0.00023	0.00275	-0.00032	<0.0001
Total Contacts	-0.00136	0.03257	0.00040	0.08116	-0.00031	0.28402
Bank report Freq	-0.01719	<0.0001	-0.00407	<0.0001	-0.01117	<0.0001
Pre recovery Rate	0.56850	<0.0001	3.63447	<0.0001	2.26212	<0.0001
EmployerNoInfo	-0.04457	0.63820	-0.01277	0.65487	0.03439	0.38375
Total Number	-0.00949	0.16151	-0.00169	0.42951	-0.00776	0.00651
γ						
(Intercept)	1.60514	<0.0001	2.64737	<0.0001	1.45450	0.00000
Pre recovery Rate	0.49096	0.00025	-2.11510	<0.0001	-0.18488	0.01538
Post Balance	0.00039	<0.0001	0.00018	NA	0.00031	0.00000
Cust Payment Freq	0.02949	<0.0001	0.17612	<0.0001	0.07759	0.00000
Credit Bureau Score	-0.00058	0.00458	-0.00033	0.09534	-0.00028	0.01388

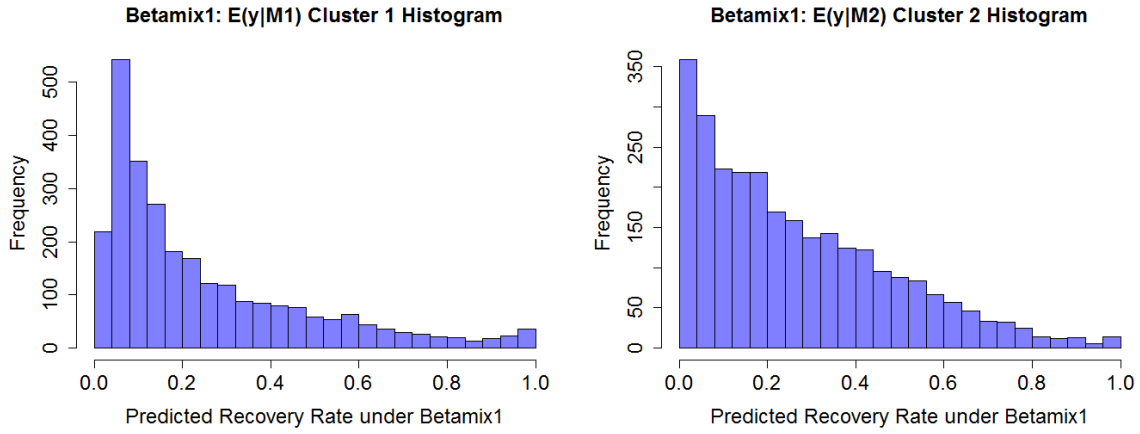
is because the estimation algorithm is unable to produce reliable standard errors in these cases. We can see that the significance of variables is *diluted* by the two clusters. For instance, credit bureau score is significant in the standard beta regression with a p-value of 0.0022, but in the beta mixture model, it is not significant for either of the clusters, taking a significance level of 5%. The direction of association of coefficient estimates in beta mixture model for both clusters are mostly consistent, where the estimates are significant (at 5% level), although magnitude of association differs. Pre-Recovery Rate for γ component is the only exception to this observation. The model also demonstrates some interesting significant associations between some variables and RR: taking insurance shows higher recoveries and having a record at the bad debt bureau is associated with lower recovery rates. Also, the recoveries, pre-purchase, are positively correlated with future RR, although total number of calls to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 February 2019 doi:10.20944/preprints201902.0135.v1

Peer-reviewed version available at Risks 2019, 7, 19; doi:10.3390/risks7010019

customer had a negative association, perhaps because these were difficult customers from whom to collect, hence requiring more intervention.

Following the procedure in Figure 5, the expected value of RR conditional on Cluster M_j is calculated based on the parameters η and γ estimated in Table 2. Since it is too time consuming to perform kernel density estimation on 29 variables, we reduce the dimension to 6 by employing PCA analysis, which greatly shortens the running time for 2 clusters' density estimations. Nevertheless, it is inevitable that information has been lost during the dimension reduction process, which may result in weaker estimates. Figure 6 shows histograms of expected value of RR conditional on each jth cluster for the test data set. The shapes of the two clusters are similar, except cluster 2 has more estimates in the range 0.2 to 0.6.



(a) $E(y|M_1, X_{new})$ **(b)** $E(y|M_2, X_{new})$

Figure 6. $E(y|M_j, X_{new})$ based on the Test data set, for the two clusters ($m = 2$

Figure 7 shows four histograms of predicted RR corresponding to the four different priors defined in Table 1, in contrast to the true RR. The predicted value of beta mixture model combined with logistic regression model is calculated by applying the formula derived in Equation 14. Models with the different priors perform in a similar way. Importantly, they are all able to model the bimodal nature of the RR. The figure shows that none of the models are good at predicting the extreme values of RR close to 0 or 1, but this natually follows from the fact that these predictions areestimates of *expected values* of RR, through Equation 12, albeit conditional on predictor variables, and so will not represent the extremes in the distribution well. Further detail can be seen in Figure 8 which shows predicted RR against true RR. The strong correlation between predicted and true RR is clear. However, it is noticeable that when true RR is around 0.6, the model tends to under-estimate for some observations. This is because the model is not perfect at detecting observations in cluster 2. This suggests future improvements to the model to enhance its capacity to predict the correct latent cluster.

4.1. Model Performance

Predictive performance is measured using K-fold cross validation with three performance measures popular in the literature on RR estimation: mean squared error (MSE), mean absolute error (MAE) and mean aggregate absolute error (MAAE). Since the sample size (8237) is relatively large and model estimation time is long, $K = 3$ is chosen. Let n be the sample size. Then,

$$MSE = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n/K}\sum_{i=1}^{n/K}(\hat{y}_{ki} - y_i)^2 = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n/K}(\hat{y}_{ki} - y_i)^2,$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 February 2019 doi:10.20944/preprints201902.0135.v1

Peer-reviewed version available at Risks 2019, 7, 19; doi:10.3390/risks7010019

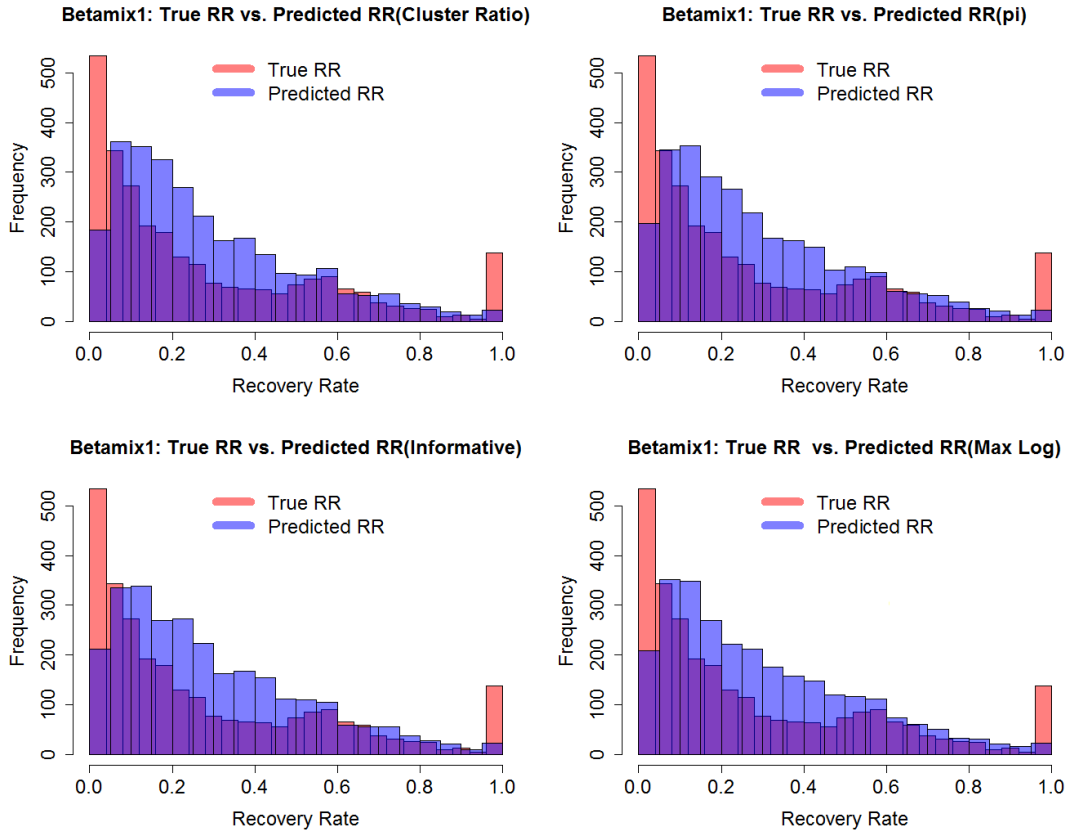


Figure 7. Predicted RR on test data ($n = 2746$) using beta mixture with four different priors, combined with logistic regression

$$MAE = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n/K}\sum_{i=1}^{n/K}|\hat{y}_{ki} - y_i| = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n/K}|\hat{y}_{ki} - y_i|.$$

MAAE is the MAE at segment level [22] and defined here as

$$MAAE = \frac{1}{V}\sum_{i=1}^{V}\frac{1}{|S_i|}\left|\sum_{j\in S_i}(y_j - \hat{y}_j)\right|$$

where V is the number of segments expressed as disjoint index sets $S_1, \cdots, S_V$. The segments could express different characteristics, eg risk bands. However, for this study each segment is a different random sample from the test data, each with approximately the same sample size and jointly exhaustive. We use $V = 100$ since this gives a balance of number of segments approximately equal to number of observations in each segment. MSE, MAE and MAAE are all penalty meaures, so the smaller the value the better the model. Since the RR is a financial ratio between 0 and 1, the MAE can reflect the size of the error in a more intuitive and direct way. If one is interested in the segment portfolio level, then the MAAE should be used.

All the models were trained on the same partitions of data into cross validation folds, to avoid bias being introduced due to different samples. Table 3 shows the results. There is little difference between results for the various linear regressions, with or without variable selection or Lasso penalty, in terms of predictive performance. The last linear model, 'excl. data set 2' was built without predictor variables from data set 2. This shows noticably worse performance than the other linear models, especially for MAE, which demonstrates that including past recoveries data (ie data set 2) gives an

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 February 2019 doi:10.20944/preprints201902.0135.v1

Peer-reviewed version available at Risks 2019, 7, 19; doi:10.3390/risks7010019

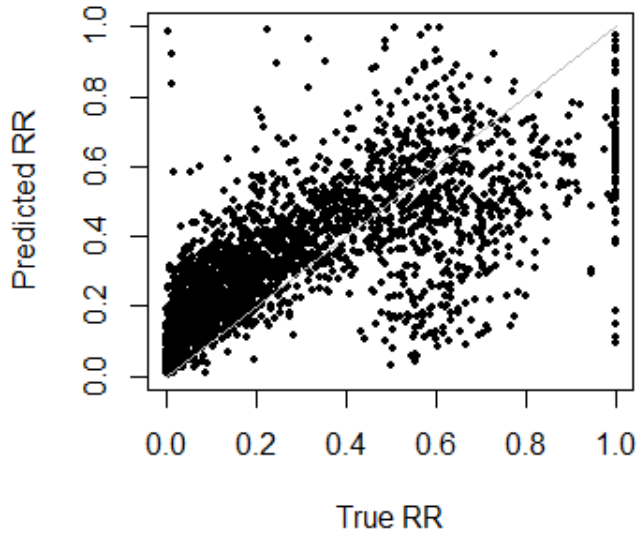


Figure 8. Predicted RR against true RR on test data ($n = 2746$) using beta mixture with the indifferent prior, combined with logistic regression

uplift in performance. The standard beta regression model, with and without zero-inflation, performs much worse than linear regression, but the beta mixture model with logistic regression gives the best performance on all three measures. The different priors give slightly different performances but are not very much different, although the Approach 1 method for selecting cluster assignment (max log-likelihood) is slightly worse than the Approach 2 soft clustering methods.

Table 3. Predictive results using 3-fold cross validation

Model	MSE	MAE	MAAE
Linear Regression			
Linear regression	0.024984	0.114268	0.025894
Stepwise linear regression	0.024752	0.113621	0.025700
Linear regression with Lasso	0.025228	0.114847	0.023739
Linear regression, excl. data set 2	0.026822	0.121385	0.026303
Beta regression			
Standard beta regression	0.085630	0.260459	0.161366
Inflated beta regression	0.076650	0.216374	0.048466
Beta mixture model combined with logistic regression			
Max log-likelihood	0.018750	0.095432	0.030629
Prior based on R Flexmix π_j	0.018460	**0.091833**	0.023991
Prior based on training set cluster size ratio	0.019325	0.092225	**0.022594**
Indifferent Prior	**0.018030**	0.092399	0.026298

5. Conclusion

Linear regression, beta regression, inflated beta regression, and a beta mixture model combined with logistic regression have been applied to model the recovery rate of non-performing loans. The models' predictive performance is measured using mean squared error, mean absolute error and mean aggregate absolute error under 3-fold cross validation. In order to produce predictions from the beta mixture model, methods of hard and soft clustering were developed and the soft clustering approaches give marginally better predictive performance. Theoretically, the proposed model, beta mixture model combined with logistic regression model, should be a suitable model to predict recovery rate for this data since it allows us to model the multimodality in the data set and takes extra consideration of the boundary value. Indeed, we found that it achieves the best results amongst the models. Stepwise linear regression also achieves relatively good performance, however, the normality and homoscedasticity assumptions do not hold. In our experiments, we also found that inclusion of previous collections data boosted predictive performance.

We believe the beta mixture model is useful for modelling RR because it is explaining different servicing strategies. In the case of our study, the cluster with mode around 0.55, is likely expressing those loans for which the debt servicer has agreed with the borrower to repay just a proportion of the outstanding debt. There may be servicing strategies in other bad debt portfolios that could be discovered using a similar mixture model or clustering approach. We have developed a technique to predict the correct latent cluster for new observations and this works well. However, results suggest that further work to refine this aspect of the use of the model could yield improved performance.

Author Contributions: conceptualization, H.Y. and T.B.; methodology, H.Y. and T.B.; validation, T.B.; investigation and statistical modelling, H.Y.; data analysis, H.Y.; writing—original draft preparation, H.Y..; writing—review and editing,T.B.; supervision, T.B.

Funding: This research received no external funding.

Acknowledgments: We wish to thank the anonymous debt collection company for use of their data and their expertise which was essential to understand the meaning and context of the data. We would also like to thank Tommaso Pappagallo who did preliminary data analysis as part of his MSci project which was useful in taking this project work forward.

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

Abbreviations

The following abbreviations are used in this manuscript:

IRB	Internal ratings based
RR	Recovery rate
LGD	Loss given default
PD	Probability of default
EAD	Exposure at default
EM	Expectation-Maximization (algorithm)
MSE	Mean square error
MAE	Mean absolute error
MAAE	Mean absolute aggregate error

Appendix

Table A1. Descriptive statistics. $n = 8237$. For numeric variables: min, mean (standard deviation), max. For factors, frequency (%age) for each level. All predictive data is collected prior to servicing.

Variable	Type	Description	Statistics
RR post	numeric	Recovery rate (outcome variable)	0.000508, 0.280 (0.283), 1
Product	factor	Type of loan	C:7468 (90.7%), R:769 (9.3%)
Principal	numeric	Original loan amount	0, 3120 (2330), 15000
Interest	numeric	Interest payments	0, 551 (439), 3380
Insurance	numeric	Insurance fees	0, 42 (84.6), 953
Late charges	numeric	Late charge fees	0, 269 (109), 1470
Overlimit fees	numeric	Over credit limit fees	0, 13.3 (24.6), 315
Creditlimit	numeric	Credit limit	0, 4560 (2660), 13800
Sex	factor	Sex	F:3196 (38.8%), M:5041 (61.2%)
Married	factor	Marriage status	0:1201 (14.6%), D:518 (6.3%), M:3929 (47.7%), O:217 (2.6%), S:2230 (27.1%), W:142 (1.7%)
Age	numeric	Age	1, 48.7 (11.1), 87
DelphiScore	integer	Credit bureau score	0, 298 (138), 443
Bureau Sub 1	factor	Loan is in the servicer's bureau (1=True)	- :1520 (18.5%), 1 :6717 (81.5%)
CustPaymentFreq	integer	Customer repayment frequency	1, 7.56 (5.59), 29
Post Balance	numeric	Exposure amount at start of servicing	0, 3130 (2630), 15900
Total paid amount	numeric	Total net paid amount	-275, 1200 (1100), 11200
Total calls	numeric	Total number of calls	0, 104 (106), 911
Total contacts	numeric	Total number of contacts (except calls)	0, 28.5 (26.5), 196
Bankreport Freq	numeric	Bank reporting frequency	0, 11.6 (7.92), 26
Pre recovery rate	numeric	Recovery rate	-0.130, 0.258 (0.217), 2.89
Employer	factor	Employer known	EmployerProvided:8053 (97.8%), NoInfo:184 (2.2%)
Total number	integer	Total number of loan accounts	0, 2.3 (2.43), 68

References

1. Bank for International Settlements. The Internal Ratings-Based Approach. Technical report, 2001.
2. Bellotti, T.; Crook, J. Loss given default models incorporating macroeconomic variables for credit cards. *International Journal of Forecasting* **2012**, *28*, 171–182.
3. Calabrese, R. Predicting Bank Loan Recovery Rates with a Mixed Continuous-Discrete Model. *Applied Stochastic Models in Business and Industry* **2012**, *30*, 99–114.
4. Papke, L.; Wooldridge, J. Econometric Methods for Fractional Response Variables With an Application to 401(K) Plan Participation Rates. *Journal of Applied Econometrics* **1996**, *11*, 619–32.
5. Qi, M.; Zhao, X. Comparison of modeling methods for Loss Given Default. *Journal of Banking & Finance* **2011**, *35*, 2842–2855.
6. Loterman, G.; Brown, I.; Martens, D.; Mues, C.; Baesens, B. Benchmarking regression algorithms for loss given default modeling. *International Journal of Forecasting* **2012**, *28*, 161–170.
7. Ji, Y.; Wu, C.; Liu, P.; Wang, J.; Coombes, K.R. Applications of beta-mixture models in bioinformatics. *Bioinformatics* **2005**, *21*, 2118–2122, [/oup/backfile/content$_p$ublic/journal/bioinformatics/21/9/10.1093/bioinformatics/bti318/2/bti318.pdf]. doi:10.1093/bioinformatics/bti318.

8. Laurila, K.; Oster, B.; Andersen, C.L.; Lamy, P.; Orntoft, T.; Yli-Harja, O.; Wiuf, C. A Beta-mixture model for dimensionality reduction, sample classification and analysis. *BMC Bioinformatics* **2011**, *12*, 215. doi:10.1186/1471-2105-12-215.
9. Moustafa, N.; Creech, G.; Slay, J. Anomaly Detection System Using Beta Mixture Models and Outlier Detection. In *Progress in Computing, Analytics and Networking. Advances in Intelligent Systems and Computing*; Pattnaik, P.; Rautaray, S.; Das, H.; Nayak, J., Eds.; Springer, Singapore, 2018; Vol. 710. doi:10.1007/978-981-10-7871-2_13.
10. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **2010**, *33*, 1–22.
11. Hastie, T.; Efron, B. *lars: Least Angle Regression, Lasso and Forward Stagewise*, 2013. R package version 1.2.
12. Ferrari, S.; Cribari-Neto, F. Beta Regression for Modelling Rates and Proportions. *Journal of Applied Statistics* **2004**, *31*, 799–815.
13. Cribari-Neto, F.; Zeileis, A. Beta Regression in R. *Journal of Statistical Software* **2010**, *34*, 1–24.
14. Nocedal, J.; Wright, S. *Numerical Optimization*, first ed.; Springer-Verlag, 1999.
15. Mittelhammer, R.; Judge, G.; Miller, D. *Econometric Foundations*, first ed.; Cambridge University Press, 2000.
16. Leisch, F. FlexMix: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software* **2004**, *11*, 1–38.
17. Gruen, B.; Leisch, F. Fitting finite mixtures of generalized linear regressions in R. *Computational Statistics & Data Analysis* **2007**, *51*, 5247–5252.
18. Gruen, B.; Leisch, F. FlexMix Version 2: Finite Mixtures with Concomitant Variables and Varying and Constant Parameters. *Journal of Statistical Software* **2008**, *28*, 1–35.
19. Gruen, B.; Kosmidis, I.; Zeileis, A. Extended Beta Regression in R: Shaken, Stirred, Mixed, and Partitioned. *Journal of Statistical Software* **2012**, *48*, 1–25.
20. Fraley, C.; Raftery., A.E. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* **2002**, *97*, 611–631.
21. Azzalini, A.; Menardi, G. Clustering via Nonparametric Density Estimation: The R Package pdfCluster. *Journal of Statistical Software* **2014**, *57*, 1–26.
22. Thomas, L.; Bijak, K. Impact of segmentation on the performance measures of LGD models. https://crc.business-school.ed.ac.uk/wp-content/uploads/sites/55/2017/02/Impact-of-Segmentation-on-the-Performance-Measures-of-LGD-Models-Lyn-Thomas-and-Katarzyna-Bijak.pdf, 2015.