Enhancing two-stage modelling methodology for loss given default with support vector machines

Enhancing Two-Stage Modelling Methodology for Loss Given Default with Support Vector Machines

Xiao Yao [a,*], Jonathan Crook [b], Galina Andreeva [c]

[a, b, c] Credit Research Centre, The University of Edinburgh Business School, 29 Buccleuch Place, Edinburgh EH8 9JS UK

Abstract: We propose to incorporate least squares support vector machine technique into a two-stage modelling framework to predict recovery rates of credit cards from a UK retail bank. The two-stage model requires a classification step that discriminates the cases with recovery rate equal to 0 or 1 and a regression step to estimate recovery rates for the cases with recovery rates in (0, 1). The two-stage model with a support vector machine classifier is found to be advantageous on an out-of-time sample compared with other methods, suggesting that a support vector machine is preferred to a logistic regression as the classification technique. We further examine the predictive performances on a subset where recovery rate is bounded in (0, 1) and the empirical evidence demonstrates that support vector regression yields significant but modest improvement compared with other statistical regression models. When modelling on the whole sample, the support vector regression does not present any advantage compared with other techniques within the two-stage modelling framework. We suggest that the choice of regression models is less influential in prediction of recovery rates than the choice of classification methods in the first step of two-stage models.

Keywords: Risk analysis, Loss given default modelling, Two-stage model, Support vector machine

1. Introduction

The Basel Accords require banks to develop their internal credit risk models for expected loss that is defined as

$$Expected\ Loss = PD \times LGD \times EAD,$$

where probability of default (PD), loss given default (LGD) and exposure at default (EAD) are the key risk parameters to be estimated in the advanced internal rating based (AIRB) approach (Basel Committee, 2005a, 2005b). In Basel II an asymptotic single risk factor model has been established to estimate PD and asset correlation proposed by Vasicek (1987) based on Merton's model (1974) with an analytical formula derived for the loss distribution under the infinite granular assumption for a given portfolio. However, under the Foundation Internal Rating Based (FIRB) approach LGD and EAD are values specified by regulators for different types of credit products. Therefore financial institutions are encouraged to develop their internal LGD model according to the requirements of the Advanced Internal Rating Based (AIRB) approach to reduce the amount of

* Corresponding author, E-mail : yaoxiao18@gmail.com

regulatory capital they are required to hold.

Together with PD modelling, LGD modelling has also arisen as a heated topic in quantitative credit risk management where extensive research has been conducted for both corporate bonds and bank retail loans. A major problem in constructing models to predict LGD is the common occurrence that the distribution of LGD values is bimodal with modes at 0 and 1. Thus two related modelling issues arise: how to transform covariates and which modelling algorithm to use.

Parametric models have been widely applied to predict the LGD of retail loans and focus on identifying potential significantly useful predictors. Qi and Yang (2009) found that loan-to-value (LTV) was useful to segment risk, but updated loan-to-value (CLTV) was the single most important determinant of modelling residential mortgage LGD. Leow and Mues (2011) developed a probability of repossession model with three variables and showed that this model performed significantly better than the model with LTV at default alone. Khieu et al (2012) examined the determinants of bank loans recovery rates by applying both OLS and fractional response regression models. They found that loan characteristics were more significant than the borrower characteristics, and that macroeconomic variables also played a significant role. This finding was consistent with the conclusions in Qi and Yang (2009). Bellotti and Crook (2012) discussed the influences of application and macroeconomic variables, and they found that the inclusion of the interaction terms of application and macroeconomic variables did not necessarily lead to an improvement of model fit. They also proposed a two-stage model to handle the bimodal distribution of recovery rates for retail credit cards. This two-stage model framework was based on a decision tree algorithm and applied to split the whole sample into three groups according to the values of recovery rates: RR=0, 0<RR<1 and RR=1. Here the extreme cases with RR=0 and RR=1 were separated by two substages such that RR=0 vs. RR>0 and RR=1 vs. 0<RR<1, and then the values in (0, 1) were fitted by an OLS regression model. The two-stage model in Bellotti and Crook (2012) showed rather robust predictive performances and outperformed many other complex models including Tobit and fractional response regression models. But the inclusion of macroeconomic variables only made a modest improvement in model fit. Bijak and Thomas (2015) proposed a Bayesian method which assumed that the LGD followed a mixture normal distribution with the weighted probability of loss following a Bernoulli distribution. This approach was able to simulate the bimodal distribution of LGD and was free of the problems discussed above. The model was estimated by a Markov Chain Monte Carlo (MCMC) procedure and applied to predicting LGD of retail unsecured loans of a UK bank. They found the estimates of the Bayesian model were very close to that estimated by the frequentist approach, and the predictive performances were also very close. Leow et al (2014) examined the effects of macroeconomic variables on two types of retail loans: residential mortgage loans and unsecured personal loans. For the mortgage loans the incorporation of economic indicators improved the model fit slightly and the LGD predictions for the loans during the economic downturn were better than for other periods, implying that macroeconomic conditions may be related to LGD non-linearly. However,

for personal loans most of the macroeconomic variables were not significant statistically and they brought no benefit for LGD predictions. They suggested the unsecured personal loans might be less affected than the mortgage loans by macroeconomic conditions.

Other parametric distributions have also been applied to LGD modelling. Calabrese (2014a) applied inflated beta regression to modelling retail loans recovery rates from the Bank of Italy. This study showed the major advantage of inflated beta regression was that it is able to analyze the different influences of the same covariates on the extreme values of 0 or 1 and the recovery rates in the interval (0, 1). Compared with fractional response regression, inflated beta regression showed consistently better out-of-sample predictive accuracies across different forecasting periods and different sample percentages of the extreme values. Furthermore, Calabrese (2014b) proposed a mixture beta distribution to estimate downturn LGD. The model was estimated based on a portfolio of bank loans of an Italian bank, and the empirical evidence showed that this mixture distribution model was able to replicate the high concentration of the loss data and thus effectively avoided underestimating the downturn LGD. However no observable characteristics were incorporated in the model.

Semi-parametric and non-parametric models have also been employed to improve the predictive accuracies for recovery rates of bank loans which were shown to be more competitive than traditional parametric models. Calabrese and Zenga (2010) presented a non-parametric mixture beta kernel estimator which incorporates the clustered cases at boundaries to predict recovery rates of loans from the Bank of Italy. Based on Monte Carlo simulation results they showed that the proposed mixture beta kernel estimator was preferable to the original beta kernel estimator for fitting a LGD distribution. But no empirical evidence using bank data was presented in this study. Bastos (2010) showed that regression trees were a competitive method to predict bank loans recovery rates compared with fractional response regression models with either a logit or a complementary log-log link function. Zhang and Thomas (2010) investigated a group of algorithms and showed that OLS was as good as the other survival models including both semi-parametric Cox hazard models and parametric accelerated failure time models, which was consistent with the findings in Bellotti and Crook (2012) for credit cards. Tong et al (2013) proposed a zero-adjusted gamma regression model by reparameterizing the mean and dispersion parameters with additive non-parametric terms. This semi-parametric model provided a flexible structure for model interpretation and effectively avoided the black box drawback by including non-parametric splines. Loterman et al (2011) benchmarked a total of 24 methods including both statistical regression models and machine learning algorithms on six bank loans loss datasets. They conducted a comprehensive study with multiple performance metrics and found that the non-linear machine learning algorithms significantly outperformed the traditional linear models. They proposed a hybrid model that combines linear and non-linear techniques and showed that it gave rather competitive predictive power while preserving the explanatory power of linear models. Tobback et al (2014) also studied the LGD of bank retail loans from two US datasets and

compared the performances of a collection of linear and non-linear models including linear regression, regression tree, support vector regression and a hybrid model similar to Loterman et al (2011). Different from the results in Loterman et al (2011), they found the best out-of-time performances were reported for the hybrid model combining a linear regression with a support vector regression on the error terms, and a regression tree showed the best out-of-sample forecasting performance for a consumer loan. They also documented the importance of incorporating macroeconomic variables, which improved the predictive performances and confirmed the impacts of business cycle on LGD that has been found in previous studies. Hwang et al (2016) proposed a similar two-stage probit model where an ordered probit model was used to predict recovery rate allocated into three categories including zero, one, and between zero and one, and a probit transformation regression was applied to estimating cases for the intermediate cases. Sun and Jin (2016) found an ensemble regression tree gave more accurate predictions than a simple regression tree or random forest but used a measure of discriminative power rathe than predictive accuracy to assess performance. Siao et al (2016) proposed a logistic quantile regression model and apply it to corporate bonds. Yao et al (2015) improved the least squares support vector regression (LS-SVR) model to account for the seniority heterogeneity and found the improved LS-SVR model outperformed the original SVR techniques and the traditional LGD regression models such as fractional response regression and linear regression. However, they did not consider a two-stage modelling method to estimate the zero or one recovery rate cases separately from the remaining cases.

Two-stage methods developed in the literature have the advantage that they address a serious problem in recovery rates modelling, which is how to model the extreme cases concentrating on the boundaries at 0 and 1. Single-stage models assume that all cases are generated from the same distribution while two-stage models have the advantage that they consider that the cases with recovery rates of 0 and 1 may be intrinsically distinct from the cases between 0 and 1 which should be separated first. We develop the hypothesis that the performances of two-stage model are disappointing owing to the probabilities generated from a logistic regression model are not accurate enough to separate the cases at boundaries from that in the interval (0, 1).

In this paper we make three contributions. First, we propose a two-stage model where 0 and 1 values are predicted using a support vector machine (SVM) classifier rather than a logistic regression model, with OLS to model intermediate values. Second, we show that this method gives more accurate predictions than other models using a large unique credit card data set. Third we find that that it is this innovation rather than the choice of algorithm to model the intermediate values that results in greater predictive accuracy. This distinguishes our work from papers that consider only one-stage models and those that consider two-stage models, the latter being Bellotti and Crook (2012) who used either logistic regression as the first stage classifier.

We seek to apply a least squares support vector classifier (LS-SVC) technique proposed by Suykens et al (1999, 2002) as an alternative method for the classification problem under the

two-stage modelling framework, and then the LS-SVC classification scores are transformed into probabilities by fitting a sigmoid form function using a maximum likelihood method proposed in Platt (1999). We choose LS-SVC for two reasons: First unlike the original SVC model proposed in Vapnik (1995, 1998), LS-SVC is more attractive for its low computational cost as it is equivalent to solving a linear system of equations instead of solving a quadratic programming problem as in SVC. Second LS-SVC was found to be consistently predictive in classifying good and bad payers on eight real-life credit scoring data sets in Baesens et al (2004) although they noted that other simple classifiers such as logistic regression also gave good performance. We consider the two-stage model in Bellotti and Crook (2012) as a benchmark for the purpose of comparison where a logistic regression was applied. We find that the two-stage model equipped with a LS-SVC method gives significantly improved predictive accuracy of recovery rates compared with the other single-stage models, which suggests that the two-stage model predictive performances rely on the choice of classification model. To further examine our hypothesis we compare the classification accuracies between LS-SVC and logistic regression methods and find that LS-SVC consistently outperforms logistic regression for both of the two substages. Finally we study how the regression method influences the two-stage framework by modelling on cases with recovery rates in [0, 1] and (0, 1) separately. We find that when modelling on the cases in (0, 1), the LS-SVR gives relatively close performances to an OLS model. But when LS-SVR is applied in the two-stage model, it is shown that the combination of LS-SVC and LS-SVR is significantly outperformed by the combination of LS-SVC and OLS statistically, although the margin is not remarkable. We conclude that the choice of regression methods plays a less crucial role than that for the classification methods.

The rest of this paper is organized as follows. Section 6.2 introduces the methodologies applied in the empirical study where the kernel based support vector machine techniques will be presented with more details. Empirical evidence will be demonstrated in Section 6.3 including the interpretations of parameters and discussions of the model performances, and Section 6.4 concludes this chapter.

2. Models

2.1 Parametric models

We study three parametric models that are commonly applied in LGD modelling including ordinary linear regression (OLS), fractional response regression and inflated beta regression methods. Both OLS and fractional response regression (Papke and Wooldridge, 1996) have been investigated extensively in LGD/recovery rates modelling for both corporate bonds and bank loans. Beta regression was proposed by Ferrari and Neto (2004) to fit the fractional response data with a beta distribution defined in (0, 1). The model is given as

$$f(y;m,f) = \frac{\Gamma(f)}{\Gamma(mf)\Gamma((1-m)f)} y^{mf-1}(1-y)^{(1-m)f-1}, \quad (1)$$

where m and f are the mean and precision parameters that can be reparameterized with respect to the predictors. However, the beta regression model defines the dependent variable y in (0, 1) and thus neglects the boundary values 0 and 1 which are especially crucial to recovery rates modelling. To overcome this drawback Ospina and Ferrari (2010) proposed an inflated beta regression model to take the boundary values into consideration. It defines a mixture distribution for the dependent variable as a combination of a Bernoulli distribution and a beta distribution such that

$$bi_{01}(y;p,\gamma,m,f) = \begin{cases} p(1-\gamma) & \text{if } y = 0 \\ p\gamma & \text{if } y = 1 \\ (1-p)f(y;m,f) & \text{if } y \in (0,1) \end{cases}. \quad (2)$$

The beta distribution assumption for recovery rates is first introduced by Gupton and Stein (2002) in Moody's internal LGD modelling framework LossCalc^{TM}. They suggested using a beta distribution to transform the recovery rates into a normally distributed space and then to employ OLS to fit the transformed dependent variable, and finally the fitted dependent variables were transformed back to the fitted recovery rates. This idea has been widely accepted and adopted in the research on LGD/recovery rates modelling (Loterman et al, 2011; Bellotti and Crook, 2012). In contrast, Calabrese (2014a) empirically studied the recovery rates of bank loans of the Bank of Italy showing that the inflated beta regression model demonstrated better out-of-time predictive accuracies compared with fractional response regression models, and that it was preferable for different forecasting periods of time and for different sample percentages of the extreme values of recovery rates. In our following study we adopt the same methodology from Calabrese (2014a) to examine whether the inflated beta regression remains an advantage on our data.

2.2. Support vector machine

The support vector machine was proposed by Vapnik (1995, 1998) and it has been increasingly attractive technique in multiple areas. Compared to other statistical models support vector models have an edge on solving non-linear problems due to the application of kernel functions. The idea of the kernel function is to map the data vectors from a low-dimension space to a high-dimension space where it is not necessary to represent the mapping function explicitly. This allows support vector models to transform a non-linear problem into a very high dimensional linear problem that gives more accurate predictions. Based on that the more recent proposed least squares support vector methods in Suykens and Vandewalle (1999) and Suykens et al (2002) have shown more excellent prediction performance with low computational cost. In this section we introduce least squares support vector methods for both classification and regression problems respectively.

Classification

Suykens and Vandewalle (1999) developed a least squares support vector classifier (LS-SVC) where the cost function was defined as the sum of squared error terms. One of the advantages of a LS-SVC is that it only needs to solve a linear system of equations instead of a quadratic

programming problem as in the standard SVM models. Given a dataset $D = \{(\mathbf{x}_i, s_i)\}_{i=1}^{N}$, $\mathbf{x}_i \in R^m$ denote the covariates of i-th observation with the related labels s_i defined as $s_i \in \{-1, 1\}$. The LS-SVC is given as

$$\begin{aligned} &\min J(\mathbf{w}, b; x_i) = \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{2}\sum_{i=1}^{N} x_i^2 \\ &s.t.\ \ s_i(\mathbf{w}^T \varphi(\mathbf{x}_i) + b) = 1 - x_i, \quad i = 1,\ldots,N \end{aligned}, \tag{3}$$

where $\mathbf{w}$ denotes the parameter vector of the associated covariates and b is the intercept term. Here error terms, x_i^2, are scaled by a regularized parameter C, and $\varphi(\mathbf{x}_i)$ represents the kernel function that maps the data from original data space to a higher dimensional space. This model is then solved by its dual form problem derived from a Lagrangian function

$$L(a_i; \mathbf{w}, b, x_i) = J(\mathbf{w}, x_i) - \sum_i a_i(s_i(\mathbf{w}^T \varphi(\mathbf{x}_i) + b) - 1 + x_i),$$

where a_i is the Lagrangian multiplier. Based on Karush-Kuhn-Tucker conditions (KKT) which are the first order sufficient conditions for an optimal solution of a non-linear mathematical programming where there is a non-linear objective function subject to a number of constraints (Boyd and Vandenberghe, 2004), we have the following equations such that

$$\begin{cases} \partial_w L = \mathbf{w} - \sum_i a_i s_i \varphi(\mathbf{x}_i) = 0 \Rightarrow \mathbf{w} = \sum_i a_i s_i \varphi(\mathbf{x}_i) \\ \partial_b L = \sum_i a_i s_i = 0 \\ \partial_{x_i} L = a_i - C x_i = 0 \Rightarrow x_i = \dfrac{a_i}{C} \end{cases}. \tag{4}$$

After inserting the optimal conditions (4) back into the Lagrangian function, a linear system of equations is formulated as follows

$$\begin{bmatrix} 0 & \mathbf{s}^T \\ \mathbf{s} & \bar{\mathbf{H}} \end{bmatrix} \begin{bmatrix} b \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{e} \end{bmatrix}, \tag{5}$$

where $\mathbf{e} = (1, \ldots, 1)^T$ (N ones), $\mathbf{s} = (s_1, \ldots, s_N)^T$, $\boldsymbol{\alpha} = (a_1, \ldots, a_N)^T$, $\bar{\mathbf{H}} = \mathbf{H} + \frac{1}{C}\mathbf{I}$, $\mathbf{H}_{ij} = s_i s_j \mathbf{K}(x_i, x_j)$, and $\mathbf{K}(x_i, x_j)$ defines the inner product of a pair of kernel functions as $\mathbf{K}(x_i, x_j) = \varphi(\mathbf{x}_i) \times \varphi(\mathbf{x}_j)$. Notice that the use of kernel functions allows the use of the inner product of the mapping function following Mercer's theorem (Mercer, 1909), without stating the mapping function explicitly, which greatly simplifies the computation.

Denote the fitted classifier as $\hat{f}$ and its predicted output as $\hat{f}(\mathbf{x}_i)$. In the following we use $\hat{f}_i$ for short. To map SVM outputs to predicted probabilities Platt (1999) proposed a parametric model to fit $\hat{f}_i$ using a sigmoid distribution, and the posterior probabilistic output $P(s_i = 1 \mid \hat{f}_i)$ is given such that

$$P(s_i = 1 \mid \hat{f}_i) = \frac{1}{1 + \exp(A\hat{f}_i + B)}, \tag{6}$$

where A and B are the unknown parameters to be estimated. The underlying assumption of this method is inspired by observing the discontinuities in the conditional densities $P(\hat{f}_i \mid s_i = \pm 1)$, and a sigmoid form function is applied to fit such discontinuities. To estimate the parameters we first redefine the target variables as

$$t_i = \frac{s_i + 1}{2},$$

and then the estimates can be obtained by minimizing the negative log likelihood of the training data iteratively defined as a cross-entropy error function such that

$$\min - \sum_i t_i \log(p_i) + (1 - t_i)\log(1 - p_i), \tag{7}$$

where $p_i = P(s_i = 1 \mid \hat{f}_i)$.

Regression

The least squares support vector regression (LS-SVR) is formulated in a similar form such that

$$\begin{aligned} &\min J(\mathbf{w}, b, u_i) = \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{2}\sum_{i=1}^{N} u_i^2 \\ &s.t.\ \ y_i = \mathbf{w}^T j(\mathbf{x}_i) + b + u_i, \quad i = 1, \ldots, N \end{aligned}, \tag{8}$$

where y_i denotes the recovery rate and we repeat the above procedure to derive the Lagrangian function such as

$$L(\mathbf{w}, b, u_i; a_i) = J(\mathbf{w}, u_i) - \sum_{i=1}^{N} a_i(\mathbf{w}^T j(\mathbf{x}_i) + b + u_i - y_i),$$

where a_i is the Lagrangian multiplier. According to the KKT conditions, the solution of the dual form is equivalent to solving the following linear equation systems

$$\begin{bmatrix} 0 & \mathbf{e}^T \\ \mathbf{e} & \bar{\mathbf{K}} \end{bmatrix} \begin{bmatrix} b \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{y} \end{bmatrix}, \tag{9}$$

where $\mathbf{e} = (1, \underbrace{\ldots, 1}_{1' N})^T$, $\mathbf{y} = (y_1, \ldots, y_N)^T$, $\boldsymbol{\alpha} = (a_1, \ldots, a_N)^T$, $\bar{\mathbf{K}} = \mathbf{K} + \frac{1}{C}\mathbf{I}$, where $\mathbf{K}$ is the kernel matrix and $\mathbf{I}$ is the identity matrix. The closed form solution is obtained as

$$\begin{cases} \boldsymbol{\alpha}^* = \bar{\mathbf{K}}^{-1}(\mathbf{y} - b^*\mathbf{e}) \\ b^* = \dfrac{\mathbf{e}^T \bar{\mathbf{K}}^{-1}\mathbf{y}}{\mathbf{e}^T \bar{\mathbf{K}}^{-1}\mathbf{e}} \end{cases}. \tag{10}$$

Finally the fitted regression model is given as below

$$\hat{g}(\mathbf{x}) = \sum_i a_i^* \mathbf{K}(\mathbf{x}_i, \mathbf{x}) + b^*. \tag{11}$$

2.3. Two-stage model

We briefly introduce the two-stage modelling framework proposed by Bellotti and Crook (2012). First define the following notations such that†

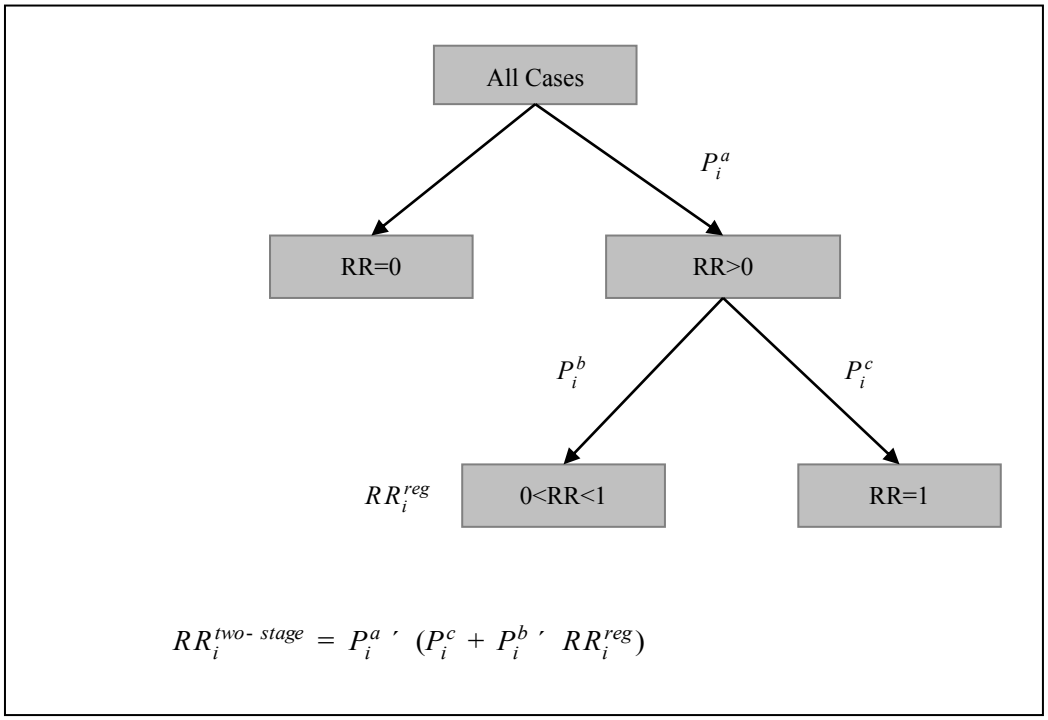
$$\begin{aligned} P_i^a &= P(RR > 0) \\ P_i^b &= P(0 < RR < 1 \mid RR > 0), \\ P_i^c &= P(RR = 1 \mid RR > 0) \end{aligned} \tag{12}$$

and then the predicted recovery rate given by a two-stage model is defined such as

$$RR_i^{two\text{-}stage} = P_i^a \times (P_i^c + P_i^b \times RR_i^{reg}), \tag{13}$$

where RR_i^{reg} denotes the predicted value by a regression model in the interval (0, 1). A diagram of two-stage modelling framework is shown in Figure 1.

Figure 1. Two-stage modelling framework



Bellotti and Crook (2012) suggested that it was normal to see that a customer in default either paid back a full proportion of the outstanding debt or paid back nothing. We believe the predictive performance of two-stage models depends on the choice of the classification methods at the final stage and thus propose to apply LS-SVC as an alternative classification method into the two-stage framework. For the regression methods we also investigate several different techniques besides OLS including fractional response regression, beta regression and LS-SVR techniques. Note that the inflated beta regression can be regarded as a hybrid model that incorporates a logistic regression and a beta regression which is analogous to a two-stage model. The difference between the two methods lies in the estimation procedure: an inflated beta regression can be estimated by solving the likelihood function in a single step and the two-stage model has to be implemented

† Please note that probabilities P_i^a, P_i^b and P_i^c may be equal to 0 or 1 although this is unusual.

step by step.

3. Empirical results

3.1. Data description and setup

A data set of credit cards used in the analysis containing recovery rates information that was provided by a major UK credit card lender. The data set consists of around 1,600,000 monthly-customer observations from March 2009 to February 2010. For the purpose of convenience the 24 months post-default recovery rate is used as the outcome variable where overdue fees and accrued interest rate are included in the recovery rate calculation[‡]. Therefore it is possible for the observed recovery rate to be less than 0 or greater than 1 for some observations. Without losing generalization we drop the cases with the recovery rates outside the range [0, 1]. Figure 2 presents a histogram of recovery rates of the whole sample. It is very clear to observe that large numbers of cases concentrate at the boundaries 0 and 1.

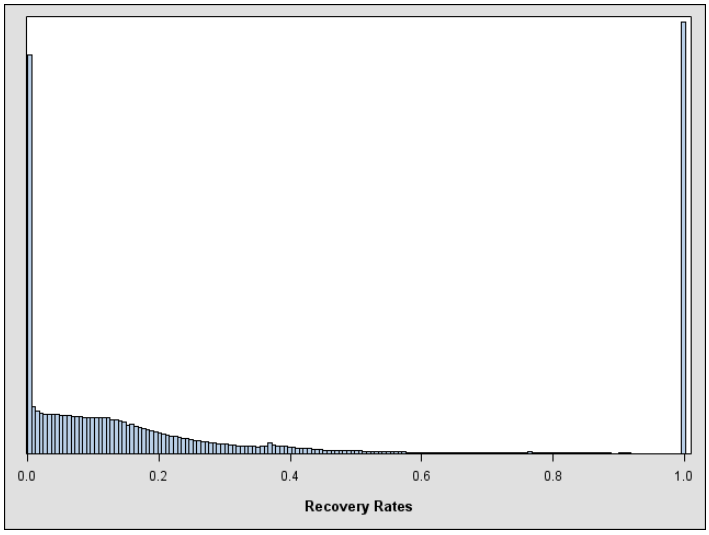
We have nearly 40 candidate predictors for recovery rates modelling recorded one year prior default. However, some of them have similar definitions and are highly correlated. We first generate the correlation matrix for all of the continuous variables and drop out the redundant ones if a correlation value is higher than 0.6 among several variables. The selected candidate variables include the account balance sheet and behavioural information. The outliers are defined as the values outside the interval between the 5 and 95 percentile of each variable and the observations with outliers are deleted. In total there are less than 5% of the total observations dropped from the total sample which will not affect the model estimates and predictions. Finally we have 13 account level variables for recovery rates models as are listed in Table 1 Panel A. Some candidate variables have been demonstrated to be important on LGD/RR modelling in literature. Bellotti and Crook (2012) showed that both *Time on Book* and *Time with Bank* had significant positive effects on recovery rate, and that *Balance at Observation* was negatively related to recovery rate. There are also some new variables that have never been investigated in literature. For example, the binary variable *Return on Order* identifies if a customer returned to order at any point in the last 12 months. It is expected to see that more outstanding debt can be recovered if the customer was shown to return. Another potentially useful predictor is whether a customer is on a repayment plan or not. It can be inferred that a customer that is on a repayment plan should have a stronger will to repay their debt than a customer that is not. Customer level information such as marital status, educational background or family income is not provided in the sample.

Macroeconomic variables are also incorporated to study the economic impacts on retail lending recovery risk. The influence of including macroeconomic variables on modelling recovery rates of unsecured retail loans is less evident than that of mortgage loans. Bellotti and Crook (2012) incorporated three variables including UK retail bank base interest rates, UK unemployment rate and UK earning index, and they found that the inclusion macroeconomic variables increased

‡ Loss given default equals to one minus recovery rate.

model fit and improves out-of-time forecasts on recovery rates, although the authors have mentioned that the data in this study spanned from 1999 to 2005 which did not cover an entire business cycle. Leow et al (2014) investigated a collection of variables from annually to quarterly and monthly indices to study the macroeconomic effects on LGD, and found that it was beneficial to incorporate macroeconomic variables for modelling the LGD of mortgage loans, but the estimates of them were almost all statistically insignificant when it comes to personal retail loans. Khieu et al (2012) explored the determinants of bank loans from Moody's database and included both economic and industry indicators, where both annual GDP growth rate and the industry distress indicator were found to affect the recovery rate significantly. Given that our data consists of monthly observations, here only monthly macroeconomic variables including UK unemployment rate, Consumer Price Index (CPI) and Housing Price Index (HPI) are included[§]. All of them are monthly data and are incorporated one month lag for each observation at default. Because of the short time period covered in our data it would not be sensible to incorporate any quarterly or yearly data. The Bank of England base interest rate is not included either because there is little change since 2008.

Figure 2. Distribution of Recovery rates[**]



The forecast accuracy of recovery rates models can be measured by the distance between the actual and predicted values, namely Root Mean Squared Errors (RMSE) and Mean Absolute Errors (MAE). R Square (R^2) is also reported that measures the proportion of the variance that is explained by the model and thus considered as an alternative performance measure fit of LGD models in literature including Qi and Zhao (2011) and Loterman et al (2011). All performance metrics are defined in (14). To test the robustness of each algorithm a cross-validation method is applied which repeatedly draws a 0.1 percent sample of the total customers randomly to create a sub-sample. The procedure is repeated for 1000 times to validate the robustness of the algorithms

[§] Macroeconomic information is sourced from Office for National Statistics (ONS). https://www.ons.gov.uk/
[**] We deliberately mask the values of Y-axis of this figure due to the confidentiality of the commercial data.

sufficiently. To assess the out-of-time predictions each sub-sample is divided into a training set from March 2009 to November 2009 and a testing set from December 2009 to February 2010. We then report the mean and standard deviations of the performance metrics for each model.

$$\begin{aligned} \text{RMSE} &= \sqrt{\frac{1}{n}\sum_i (r_i - \hat{r}_i)^2} \\ \text{MAE} &= \frac{1}{n}\sum_i |r_i - \hat{r}_i| \\ \text{R}^2 &= 1 - \frac{\sum_i (r_i - \hat{r}_i)^2}{\sum_i (r_i - \bar{r})^2}, \quad \bar{r} = \frac{1}{N}\sum_i r_i \end{aligned} \tag{14}$$

3.2. Multivariate analysis

An explanatory analysis is performed on the whole sample using a linear regression model. Here two regression models are estimated: 1) with account level variables only, 2) with both account and macroeconomic variables. The outputs of parameters estimates and model fit are reported in Panel A and B of Table 1 respectively. To show the degree of multi-colinearity the VIF values for each parameter are also reported. It should be noticed that no variable has a VIF value greater than 5, which indicates that the model estimates are not significantly affected by multi-colinearity. Table 1 shows that with the inclusion of macroeconomic variables improves R^2 modestly from 0.1508 to 0.1515 although all three macroeconomic variables are statistically significant. It is observed that all the account level variables remain significant at 0.01 confidence level with the inclusion of macroeconomic variables, indicating all account level variables are conditionally correlated with recovery rates.

Some straightforward conclusions on estimates of parameters can be taken from Table 1. For example, the number of months the account was with the bank (*Time in months*) and the number of months that the customer has held the credit card (*Time on book*) both positively influence the recovery rate, showing that the longer a customer stays with the bank, a higher proportion of its debt will be recovered after default. According to Bellotti and Crook (2012), these two variables are the indicators of customer stability which are expected to lead to a lower recovery risk. *Balance at Observation* is shown to be negatively correlated with recovery rate, which indicates that the more outstanding debt a customer has, the more difficult it is to recover. It can be observed that the longer the customer is in arrear, the more will be repaid to the bank according to Table 1. One would expect that a bank would take more actions to urge the customer to pay back its debt if it finds the customer has been in default for a long time. For the repayment behaviours it shows that the number of post-default payments made in last 12 months positively affects recovery rate, and as expected the average payment as percentage of balance also positively influences the recovery rate. It is also expected to observe a higher recovery rate if a customer makes a higher payment most recently. There are three binary variables relating to the status of recovery process. Specifically a higher recovery rate is expected if a customer returned to order in the last 12 months which is shown in Table 1. However, contrary to expectation that a customer is

on a repayment plan influences the recovery rate negatively, we suggest it is because the customer who is assumed not to be able to repay its debt may be forced to join the repayment plan and is less capable of repaying debt. It can be found that a customer that has spent 1-5 months in arrears appears to repay more debt than otherwise.

Turning to macroeconomic variables we notice that both CPI and HPI are both negatively and significantly related to recovery rate, which implies that when price inflation increases customers are less capable of paying back their outstanding debts. The puzzling sign of the estimate of unemployment rate conflicts with the finding in Bellotti and Crook (2012), where the unemployment rate was shown to be negatively correlated to recovery rate. They also found that the inclusion of macroeconomic variables generally improves the recovery rates predictions across test quarters modestly. However, our sample data spans only one year and a data set with a longer time window is needed to investigate the impacts of macroeconomic conditions on modelling unsecured loans recovery rates.

Table 1. Explanatory analysis

Table 1 shows the model estimates, variance inflation factor (VIF) and goodness-of-fit of a linear regression fitting on recovery rates on the whole sample. Panel A shows the regression outputs with account characteristics included only, and Panel B presents the outputs with both account and macroeconomic factors.

Panel A. Modelling with account variables

	Estimate	p value	VIF
Intercept	0.4586 *** (0.0030)	<.0001	0
Time in months	0.0054 *** (0.0008)	<.0001	1.1881
Time on book	0.0002 *** (0.0000)	<.0001	1.2708
Sum of transactions across all current accounts	0.0064 *** (0.0001)	<.0001	1.0342
Number of months in arrears six months ago	0.0327 *** (0.0003)	<.0001	1.5260
Balance at observation	-0.0136 *** (0.0002)	<.0001	1.1486
Worst delinquency status in days across all products	-0.0001 *** (0.0000)	<.0001	1.0397
Number of payments made last 12 months	0.0025 *** (0.0003)	<.0001	2.2370
Average payment as percentage of balance in default summed over last 6 months	8.5094 *** (0.0852)	<.0001	2.0001
Status on if a customer returned to order	0.0029 * (0.0016)	0.0719	2.1807
Status on if a customer has spent 1-5 months in arrears	0.0189 *** (0.0012)	<.0001	2.0147
Status on if a customer is on a repayment plan	-0.1664 *** (0.0019)	<.0001	1.9355
Most recent payment received	0.0052 *** (0.0011)	<. 0001	1.5109
F value	24653.0	<.0001	
R^2	0.1500		
Adj R^2	0.1500		
RMSE	0.3297		

Panel B. Modelling with account and macroeconomic variables

	Estimate	p value	VIF
Intercept	0.6255 *** (0.0595)	<.0001	0
Time in months	0.0054 *** (0.0003)	<.0001	1.1894
Time on book	0.0002 *** (0.0000)	<.0001	1.2717
Sum of transactions across all current accounts	0.0064 *** (0.0000)	<.0001	1.0355
Number of months in arrears six months ago	0.0326 *** (0.0002)	<.0001	1.5391
Balance at observation	-0.0136 *** (0.0000)	<.0001	1.1518
Worst delinquency status in days across all products	-0.0001 *** (0.0000)	<.0001	1.0474
Number of payments made last 12 months	0.0025 *** (0.0001)	<.0001	2.2429
Average payment as percentage of balance in default summed over last 6 months	8.5050 *** (0.0387)	<.0001	2.0101
Status on if a customer returned to order	0.0021 *** (0.0008)	0.0051	2.1852
Status on if a customer has spent 1-5 months in arrears	0.0192 *** (0.0008)	<.0001	2.0151
Status on if a customer is on a repayment plan	-0.1650 *** (0.0009)	<.0001	1.9469
Most recent payment received	0.0050 *** (0.0006)	0.0018	1.5120
Monthly unemployment rate	0.0899 *** (0.0033)	<.0001	1.0229
Monthly CPI	-0.0035 *** (0.0003)	<.0001	1.8726
Monthly HPI	-0.0009 *** (0.0000)	<.0001	1.8577
F value	19822.1	<.0001	
R^2	0.1507		
Adj R^2	0.1507		
RMSE	0.3296		

3.3. Out-of-sample predictions

To investigate the effects of classification and regression in two-stage models we propose to model the cases with RR in [0, 1] and (0, 1) separately. Single-stage and two-stage models are all compared in [0, 1] and only single-stage models are benchmarked in (0, 1). Four methods are

investigated to be single-stage models including OLS, fractional response regression, inflated beta regression and LS-SVR. For the two-stage models two classification methods are applied including logistic regression and LS-SVM, and there are four regression methods employed for the second stage that are the same as single-stage models except that the inflated beta regression is replaced by a beta regression model. In total we have eight combinations for the two-stage models (See Table 2). In the following the abbreviations of two-stage models names are used for convenience[††]. For example, the combination of LS-SVM and fractional response regression is abbreviated as SVM+Frac.

We first analyze the predictive performances of the cases with RR in [0, 1] and report the outputs in Table 2[‡‡]. To compare model performances the two sample t-test is applied to RMSE and MAE and both the differences between each pair of models and the p values are reported in Table 3. It should be noted that OLS outperforms the other generalized linear models including fractional response regression and inflated beta regression models in terms of out-of-sample prediction performances. Such evidence is expected although the empirical recovery rates distribution is far from a Gaussian distribution. Both Zhang and Thomas (2010) and Bellotti and Crook (2012) have reported that the OLS regression model gave better predictions than other generalized linear models. Empirical evidence in Zhang and Thomas (2010) suggested that the flexibility of survival regression did not necessarily give better predictions because it was difficult to separate from the zero recovery rates cases for the accelerated failure time models. In our study inflated beta regression, which is designed to accommodate the cases at the boundaries 0 and 1, does not show any advantages compared with OLS and fractional response regression. It is clear that SVR yields better model fit and predictive accuracy for both in-sample and out-of-sample tests. This result is also consistent with the findings in Loterman et al (2011) which showed SVR and neural networks significantly outperformed the other linear models for LGD prediction implying a strong non-linear relationship between LGD and its predictors.

Performances of two-stage models are more straightforward. The two-stage logistic+OLS method proposed in Bellotti and Crook (2012) gave slightly better out-of-sample predictions than the single-stage OLS model. We replace the OLS with other techniques and find no noticeable improvement for either logistic+Frac or logistic+Beta. Instead logistic+OLS gives significantly better out-of-sample predictive accuracy than those. Furthermore it is noticed that logistic+SVR has significantly lower R^2 and MAE and an insignificant improvement in terms of RMSE compared with logistic+OLS according to Table 3 Panel B. It indicates that the non-linear methods are not shown to improve the performances of two-stage models.

To examine the hypothesis developed above the logistic regression model is replaced by a LS-SVC technique under the two-stage modelling framework, and it shows that the two-stage

[††] The term of reference of all model names is given in Table 7.

[‡‡] Depending on the computer specification it takes between 15 and 25 minutes to run a 1000 times cross-validation for a two-stage model logistic+OLS, for SVC+Frac it takes between 60 and 90 minutes, and for SVC+SVR it takes between 90 and 120 minutes. Please note the computation times given above are for reference only.

models SVC+OLS and SVC+Frac significantly outperform all the other models. As can be seen from Table 3 there are insignificant differences between SVC+OLS and SVC+Frac in terms of R^2 and RMSE although SVC+Frac shows a slightly significant better MAE. Note that neither SVC+Beta nor SVC+SVR show better predictive accuracies than SVC+OLS or SVC+Frac. Also it is observed that the SVC+Beta model is much less competitive than any other two-stage method with a SVC technique. But SVC+Beta significantly outperforms the other single-stage statistical models, which implies that the cases with RR in (0, 1) may have a linear relationship between recovery rate and its predictors. Combined with the consistently poor performances of the inflated beta regression model, it indicates that a beta distribution is not proving to be a superior model for recovery rates as expected. Yet when the SVC technique is applied as the classification method, all two-stage models present noticeable improvements compared with those using a logistic regression, suggesting that the probabilities of recovery rates being 0 or 1 generated in equation (12) from SVC techniques are more accurate than that from logistic regression models.

Table 2. Model performances on cases with RR in [0, 1]

Table 2 presents the out-of-sample predictive accuracy for single-stage and two-stage models respectively. All models are indexed from Model1 to Model12 as follows: Model1: Ordinary linear regression (OLS), Model2: Fractional response regression, Model3: Inflated beta regression, Model4: LS-SVR, Model5: Logistic regression+OLS, Model6: Logistic regression+Fractional response regression, Model7: Logistic regression+Beta regression, Model8: Logistic regression+LS-SVR, Model9: LS-SVC+OLS, Model10: LS-SVC+Fractional response regression, Model 11: LS-SVC+Beta regression, Model12: LS-SVC+LS-SVR.

Panel A. Single-stage models

	In sample			Out of sample		
	R^2	RMSE	MAE	R^2	RMSE	MAE
Model1	0.2014	0.2449	0.3176	0.0882	0.3424	0.2634
	(0.0393)	(0.0326)	(0.0187)	(0.0694)	(0.0213)	(0.0266)
Model2	0.2030	0.3173	0.2413	0.0778	0.3443	0.2678
	(0.0458)	(0.0329)	(0.0192)	(0.0707)	(0.0217)	(0.0273)
Model3	0.0690	0.3431	0.2721	0.0179	0.3556	0.2864
	(0.0318)	(0.0333)	(0.0252)	(0.0146)	(0.0221)	(0.0284)
Model4	0.6471	0.2112	0.1541	0.1214	0.3363	0.2538
	(0.0310)	(0.0283)	(0.0126)	(0.0570)	(0.0213)	(0.0219)

Panel B. Two-stage models

	In sample			Out of sample		
	R^2	RMSE	MAE	R^2	RMSE	MAE
Model5	0.2302	0.3118	0.2316	0.1018	0.3398	0.2549
	(0.0481)	(0.0335)	(0.0194)	(0.0723)	(0.0221)	(0.0266)
Model6	0.2194	0.3203	0.2355	0.0894	0.3417	0.2509
	(0.0542)	(0.0682)	(0.0191)	(0.0698)	(0.0220)	(0.0246)
Model7	0.2207	0.3131	0.2351	0.0880	0.3423	0.2513
	(0.0452)	(0.0164)	(0.0197)	(0.0736)	(0.0220)	(0.0251)
Model8	0.2871	0.2938	0.2099	0.0825	0.3389	0.2616
	(0.0483)	(0.0191)	(0.0219)	(0.0709)	(0.0222)	(0.0240)
Model9	0.4794	0.2553	0.1707	0.1710	0.3256	0.2534
	(0.0430)	(0.0142)	(0.0220)	(0.0607)	(0.0215)	(0.0359)
Model10	0.4771	0.2550	0.1726	0.1744	0.3263	0.2509
	(0.0493)	(0.0151)	(0.0141)	(0.0654)	(0.0182)	(0.0202)
Model11	0.4534	0.2612	0.1874	0.1329	0.3349	0.2605
	(0.0475)	(0.0137)	(0.0226)	(0.0787)	(0.0217)	(0.0375)
Model12	0.5476	0.2378	0.1483	0.1628	0.3278	0.2444
	(0.0465)	(0.0156)	(0.0209)	(0.0767)	(0.0219)	(0.0245)

Table 3. Out-of-sample comparisons on [0, 1]

Table 3 presents the absolute differences of RMSE and MAE between the model of a related row and the model of a related column. Two sample t-test is conducted with p-values reported in parenthesis. A positive t-value indicates that the model performance metric of related row is higher than the model of related column and vice versa. Symbols such as ***, ** and * indicate the significance at 0.01, 0.05 and 0.1 confidence level respectively.

All models are indexed from Model1 to Model10 as follows: Model1: OLS, Model2: Fractional response regression, Model3: Inflated beta regression, Model4: LS-SVR, Model5: Logistic regression+OLS, Model6: Logistic regression+Fractional response regression, Model7: Logistic regression+Beta regression, Model8: Logistic regression+LS-SVR, Model9: LS-SVC+OLS, Model10: LS-SVC+Fractional response regression, Model 11: LS-SVC+Beta regression, Model12: LS-SVC+LS-SVR.

Panel A. RMSE

	Model1	Model2	Model3	Model4	Model5	Model6	Model7	Model8	Model9	Model10	Model11	Model12
Model1	-											
Model2	0.0019 **	-										
	(0.0483)											
Model3	0.0132 ***	0.0113 ***	-									
	(0.0000)	(0.0000)										
Model4	-0.0061 ***	-0.0080 ***	-0.0193 ***	-								
	(0.0000)	(0.0000)	(0.0000)									
Model5	-0.0026 ***	-0.0045 ***	-0.0158 ***	0.0035 ***	-							
	(0.0075)	(0.0000)	(0.0000)	(0.0003)								
Model6	-0.0007	-0.0026 ***	-0.0139 ***	0.0054 ***	0.0019 *	-						
	(0.4698)	(0.0079)	(0.0000)	(0.0000)	(0.0542)							
Model7	-0.0001	-0.0020 **	-0.0133 ***	0.0060 ***	0.0025 **	0.0006	-					
	(0.9178)	(0.0408)	(0.0000)	(0.0000)	(0.0113)	(0.5420)						
Model8	-0.0035 ***	-0.0054 ***	-0.0167 ***	0.0026 ***	-0.0009	-0.0028 ***	-0.0034 ***	-				-
	(0.0003)	(0.0000)	(0.0000)	(0.0076)	(0.3637)	(0.0047)	(0.0006)					
Model9	-0.0168 ***	-0.0187 ***	-0.0300 ***	-0.0107 ***	-0.0142 ***	-0.0161 ***	-0.0167 ***	-0.0133 ***	-			
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)				
Model10	-0.0161 ***	-0.0180 ***	-0.0293 ***	-0.0100 ***	-0.0135 ***	-0.0154 ***	-0.0160 ***	-0.0126 ***	0.0007	-		
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.4321)			
Model11	-0.0075 ***	-0.0094 ***	-0.0207 ***	-0.0014	-0.0049 ***	-0.0068 ***	-0.0074 ***	-0.0040 ***	0.0093 ***	0.0086 ***	-	
	(0.0000)	(0.0000)	(0.0000)	(0.1456)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)		
Model12	-0.0146 ***	-0.0165 ***	-0.0278 ***	-0.0085 ***	-0.0120 ***	-0.0139 ***	-0.0145 ***	-0.0111 ***	0.0022 **	0.0015 *	-0.0071 ***	
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0235)	(0.0959)	(0.0000)	

Model1: OLS, Model2: Fractional response regression, Model3: Inflated beta regression, Model4: LS-SVR, Model5: Logistic regression+OLS, Model6: Logistic regression+Fractional response regression, Model7: Logistic regression+Beta regression, Model8: Logistic regression+LS-SVR, Model9: LS-SVC+OLS, Model10: LS-SVC+Fractional response regression, Model 11: LS-SVC+Beta regression, Model12: LS-SVC+LS-SVR.

Panel B. MAE

	Model1	Model2	Model3	Model4	Model5	Model6	Model7	Model8	Model9	Model10	Model11	Model12
Model1	-											
Model2	0.0044 ***	-										
	(0.0003)											
Model3	0.0230 ***	0.0186 ***	-									
	(0.0000)	(0.0000)										
Model4	-0.0096 ***	-0.0140 ***	-0.0326 ***	-								
	(0.0000)	(0.0000)	(0.0000)									
Model5	-0.0085 ***	-0.0129 ***	-0.0315 ***	0.0011	-							
	(0.0000)	(0.0000)	(0.0000)	(0.3128)								
Model6	-0.0125 ***	-0.0169 ***	-0.0355 ***	-0.0029 ***	-0.0040 ***	-						
	(0.0000)	(0.0000)	(0.0000)	(0.0054)	(0.0005)							
Model7	-0.0121 ***	-0.0165 ***	-0.0351 ***	-0.0025 **	-0.0036 ***	0.0004	-					
	(0.0000)	(0.0000)	(0.0000)	(0.0177)	(0.0019)	(0.7190)						
Model8	-0.0018	-0.0062 ***	-0.0248 ***	0.0078 ***	0.0067 ***	0.0107 ***	0.0103 ***	-				
	(0.1123)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)					
Model9	-0.0100 ***	-0.0144 ***	-0.0330 ***	-0.0004	-0.0015	0.0025 *	0.0021	-0.0082 ***	-			
	(0.0000)	(0.0000)	(0.0000)	(0.7636)	(0.2885)	(0.0694)	(0.1297)	(0.0000)				
Model10	-0.0125 ***	-0.0169 ***	-0.0355 ***	-0.0029 ***	-0.0040 ***	0.0000	-0.0004	-0.0107 ***	-0.0025 *	-		
	(0.0000)	(0.0000)	(0.0000)	(0.0021)	(0.0002)	(1.0000)	(0.6947)	(0.0000)	(0.0551)			
Model11	-0.0029 **	-0.0073 ***	-0.0259 ***	0.0067 ***	0.0056 ***	0.0096 ***	0.0092 ***	-0.0011	0.0071 ***	0.0096 ***	-	
	(0.0462)	(0.0000)	(0.0000)	(0.0000)	(0.0001)	(0.0000)	(0.0000)	(0.4347)	(0.0000)	(0.0000)		
Model12	-0.0190 ***	-0.0234 ***	-0.0420 ***	-0.0094 ***	-0.0105 ***	-0.0065 ***	-0.0069 ***	-0.0172 ***	-0.0090 ***	-0.0065 ***	-0.0161 ***	-
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	

We further explore the advantages of SVC techniques by comparing the classification accuracies of logistic regression and SVC models in terms of AUC (Area under curve). AUC is a statistics related to a ROC (Receiver Operating characteristic) curve to measure the overall performance of the classifier scores. A simple method of AUC calculation of a classifier G was presented in Hand and Till (2001) as equation (14).

$$\hat{AUC} = \frac{\sum_i r_i - n_0(n_0 + 1)/2}{n_0 n_1}, \qquad (14)$$

where n_0 and n_1 are the numbers of positive and negative cases respectively[§§], and r_i denotes the rank of *i*-th positive case in the ranked list of the predictive values from the logistic regression. In two stage models there are two classification events involved. Event1: RR=0 vs. RR>0; Event2: RR=1 vs. 0<RR<1. The same scheme is applied as we did for recovery rates prediction to generate the classification predictions repeatedly for 1000 times and report both the means and the standard errors of both in-sample and out-of-sample performances in Table 4 for both events respectively. A two sample t-test is employed for out-of-sample predictions comparisons. It shows that SVM models excel in general for both events in terms of in-sample and out-of-sample AUC. It is also noticed that both logistic regression and SVM perform fairly well for Event1 with an AUC higher than 0.85 with mild advantage shown by SVM. SVM gives a better performance on Event2 with significant improvement on AUC. Table 5 confirms the expectations that the SVM technique is able to generate more accurate predicted probabilities than logistic regression. It should be noted that it is more difficult to separate the cases with RR in (0, 1) from that with RR=1, suggesting that the customers who are willing to repay all debts are more difficult to be separated compared with those who are unable to repay any debt.

Table 4. AUC comparisons of classification
Event1: RR=0 vs. RR>0; Event2: RR=1 vs. 0<RR<1.

	In-sample		Out-of-sample	
	Event 1	Event 2	Event 1	Event 2
Logistic Regression	0.9089	0.8152	0.8671	0.7648
	(0.0248)	(0.0311)	(0.0406)	(0.0544)
LS-SVC	0.9245	0.9549	0.8725	0.8013
	(0.0300)	(0.0116)	(0.0443)	(0.0572)
t value			-2.84 ***	-4.61 ***
			(0.0045)	(0.0000)

According to above it could be suggested that non-linear models including both statistical and machine learning techniques do not exhibit advantages over OLS in the two-stage frameworks no matter whether a logistic regression or a SVC technique is used as the classification method. To examine the effects of regression models four methods are applied to estimating RR in (0, 1) including OLS, fractional response regression, beta regression and a SVR technique. The performance metrics and model comparison results are reported in Tables 5 and 6 respectively. According to Table 6 the SVR technique outperforms the other methods significantly in terms of out-of-sample MAE, but it presents an insignificant advantage compared with OLS in terms of R^2 and RMSE. It suggests that the SVR is considered to be as accurate as OLS when modelling RR in

[§§] In this study for the first stage the cases with RR=0 are positive and RR>0 are negative, and for the second stage the cases with RR=1 are positive and 0<RR<1 are negative.

(0, 1), and both of them are significantly better than the fractional response and beta regression models. This is consistent with the evidence presented above and it can be concluded that SVC+OLS and SVC+SVR give similarly accurate out-of-sample predictions. Similarly Logistic+OLS shows marginal advantage over Logistic+SVR in terms of R^2 and MAE.

Table 5. Performances of single-stage models in 0<RR<1

Table 5 presents the in-sample and out-of-sample performances of single-stage model on the sample with recovery rate between (0, 1). All models are indexed from Model1 to Model4 as follows: Model1: OLS, Model2: Fractional response regression, Model3: Inflated beta regression, Model4: LS-SVR.

	In sample			Out of sample		
	R^2	RMSE	MAE	R^2	RMSE	MAE
Model1	0.1951	0.1564	0.1072	0.0792	0.2037	0.1480
	(0.0567)	(0.0165)	(0.0114)	(0.0944)	(0.0270)	(0.0215)
Model2	0.1705	0.1589	0.1097	0.0513	0.2069	0.1476
	(0.0547)	(0.0166)	(0.0117)	(0.0906)	(0.0272)	(0.0205)
Model3	0.0633	0.1686	0.1180	0.0277	0.2179	0.1538
	(0.1073)	(0.0192)	(0.0161)	(0.0726)	(0.0306)	(0.0282)
Model4	0.6579	0.1015	0.0691	0.0828	0.2035	0.1363
	(0.0370)	(0.0113)	(0.0076)	(0.0827)	(0.0265)	(0.0160)

Table 6 Comparisons of single-stage models

Table 6 presents the absolute differences of RMSE and MAE between the model of related row and the model of related column. Two sample t-test is conducted with p values reported in parenthesis. A positive t-value indicates that the model performance metric of related row is higher than the model of related column and vice versa. Symbols such as ***, ** and * indicate the significance at 0.01, 0.05 and 0.1 confidence level respectively.

All models are indexed from Model1 to Model4 as follows: Model1: OLS, Model2: Fractional response regression, Model3: Inflated beta regression, Model4: LS-SVR.

Panel A. RMSE

	Model1	Model2	Model3	Model4
Model1	-			
Model2	0.0032 ***	-		
	(0.0083)			
Model3	0.0142 ***	0.0110 ***	-	
	(0.0000)	(0.0000)		
Model4	-0.0002	-0.0034 ***	-0.0144 ***	
	(0.8672)	(0.0047)	(0.0000)	

Panel B. MAE

	Model1	Model2	Model3	Model4
Model1	-			
Model2	-0.0004 (0.6703)	-		
Model3	0.0058 *** (0.0000)	0.0062*** (0.0000)	-	
Model4	-0.0117 *** (0.0000)	-0.0113 *** (0.0000)	-0.0175 *** (0.0000)	-

Table 7 Term of reference

Table 7 presents the abbreviations and full names of the models in this study.

Models	**Abbreviations**	**Full names**
Model1	OLS	Ordinary Linear Regression
Model2	Frac	Fractional Response Regression
Model3	Inflated Beta	Inflated Beta Regression
Model4	SVR	Least Squared Support Vector Regression
Model5	Logistic+OLS	Logistic Regression and Ordinary Linear Regression
Model6	Logistic+Frac	Logistic Regression and Fractional Response Regression
Model7	Logistic+Beta	Logistic Regression and Beta Regression
Model8	Logistic +SVR	Logistic Regression and Least Squared Support Vector Regression
Model9	SVC+OLS	Least Squared Support Vector Classification and Ordinary Linear Regression
Model10	SVC+Frac	Least Squared Support Vector Classification and Fractional Response Regression
Model11	SVC+Beta	Least Squared Support Vector Classification and Beta Regression
Model12	SVC+SVR	Least Squared Support Vector Classification and Least Squared Support Vector Regression

4. Conclusions

This paper evaluates the performances of a group of statistical and machine learning techniques to predict the recovery rates of a large sample of UK bank credit cards, and shows that machine learning techniques are an effective supplement to statistical regression models to improve the recovery rates predictions. The kernel based least squares support vector machine techniques are applied in two ways. First the recovery rates are modelled with support vector regression directly which demonstrates better predictions than the other linear or generalized linear models in terms of both in-sample and out-of-sample predictive metrics on average, although the improvements of RMSE and MAE are not as remarkable as R^2. Second the support vector machine is incorporated into a two-stage modelling framework where the cases with zero and one recovery rates are separated by a least squares support vector classifier, and then the cases in the interval (0, 1) are modelled with other regression models. It can be concluded that the combination of LS-SVC and OLS gives the best out-of-sample predictive accuracies in terms of out-of sample RMSE and MAE. It is also noticed that this two-stage model outperforms the single-stage support vector regression model significantly in terms of the out-of-sample R^2. For the other combinations of two-stage models, where the OLS is replaced by other statistical or machine learning methods, the predictive performances are not as good as the SVC+OLS model. We suggest that choice of algorithm at the separation stage of the two-stage model plays an evidently crucial role in the predictive accuracy of recovery rates modelling while the choice of algorithms at the regression stage in (0, 1) is less important.

References

Baesens, B., Gestel, T. V., Viaene, S., Stepanova, M., Suykens, J., & Vanthiene, J. (2004). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society* 54, 627-635.

Basel Committee on Banking Supervision (2005a). Guidance on paragraph 468 of the framework document.

Basel Committee on Banking Supervision (2005b). An explanatory note on the Basel II IRB risk weight functions.

Bastos, J. A. (2010). Forecasting bank loans loss-given-default, *Journal of Banking and Finance* 34(10), 2510-2517.

Bellotti, T. & Crook, J. (2012). Loss given default models incorporating macroeconomic variables for credit cards. *International Journal of Forecasting* 28(1), 171-182.

Bijak, K. & Thomas, L. (2015). Modelling LGD for unsecured retail loans using Bayesian methods. *Journal of Operational Research Society*, 66(2), 342-352.

Boyd, S. & Vandenberghe, L (2004), Convex Optimization, Cambridge University Press.

Calabrese, R. & Zenga, M. (2010). Bank loan recovery rates: Measuring and nonparametric density estimation. *Journal of Banking and Finance* 34, 903-911.

Calabrese, R. (2014a). Predicting bank loan recovery rates in a mixed continuous-discrete model, *Applied Stochastic Models in Business and Industry* 30(2), 99-114.

Calabrese, R. (2014b). Downturn Loss Given Default: Mixture distribution estimation. *European Journal of Operational Research* 237, 271-277.

Ferrari, S. & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics* 31(7), 799–815.

Gupton, G. M. & Stein, R. M. (2002). LossCalc^{TM}: Model for predicting loss given default (LGD). *Moody's KMV*.

Hand, D. J. & Till, R. J. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning* 45, 171-186.

Hwang, R. C., Chung, H., & Chu, C. K. (2016). A two-stage Probit model for predicting recovery rates. *Journal of Financial Services Research* 50(3) 311-339.

Khieu, H. D. Mullineaux, D. J. & Yi, H. C. (2012). The determinants of bank loan recovery rates. *Journal of Banking and Finance*, 36(4), 923-933.

Mercer, J. (1909). Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society A*, 209(441–458): 415–446.

Merton, R. C. (1974). The pricing of corporate debt: The risk structure of interest rates. *Journal of Finance*, 29, 449-470.

Leow, M. & Mues, C. (2011). Predicting loss given default (LGD) for residential mortgage loans: a two-stage model and empirical evidence for UK bank data. *International Journal of Forecasting*, 28(1), 183-195.

Leow, M., Mues, C. & Thomas, L. (2014). The economy and loss given default: evidence form two UK retail lending data sets. *Journal of Operational Research Society* 65(3), 363-375.

Loterman, G., Brown, I., Martens, D., Mues, C. & Baesens, B. (2011). Benchmarking regression algorithms for loss given default modelling. *International Journal of Forecasting* 28(1), 161-170.

Ospina, R. & Ferrari, S. (2010). Inflated beta distributions, *Statistical Papers* 51, 111-126.

Papke, L. & Wooldridge, J. (1996). Econometric method for fractional response variables with an application to the 401(K) plan participation rates. *Journal of Applied Econometrics* 11, 619-632.

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In Advances in Large Margin Classifiers, Smola, A., Bartlett, P., Scholkopf, B. & Schuurmans, D. MIT Press.

Qi, M. & Yang, X. (2009). Loss given default of high loan-to value residential mortgages. *Journal of Banking and Finance* 33, 788-799.

Qi, M. & Zhao, X., (2011). Comparison of modeling methods for loss given default. *Journal of Banking and Finance* 35, 2842-2855.

Siao, J-S. S., Hwang, R-C. & Chu, C-K. (2016). Predicting recovery rates using logistic quantile regression with bounded outcomes. *Quantitative Finance* 16(5), 777-792.

Sun, H S & Jin, Z. (2016). Estimating credit risk parameters using ensemble learning methods: an empirical study on loss given default. *Journal of Credit Risk* 12(3), 43-69.

Suykens, J. & Vandewalle, J. (1999). Least Squares Support Vector Machine Classifiers. *Neural Processing Letters* 9, 293-300.

Suykens, J., Gestel, T. Van, Brabanter, J. De, Moor, B. De & Vandewalle, J. (2002). Least Squares Support Vector Machine, World Scientific, Singapore.

Tobback, E., Martens, David., Gestel, T & Baesens, B. (2014). Forecasting loss given default models: Impact of account characteristics and the macroeconomic state. *Journal of Operational Research Society* 65(3), 376-392.

Tong, E. N. C., Mues, C. & Thomas, L. (2013). A zero-adjusted gamma model for mortgage loan loss given default. *International Journal of Forecasting* 29(4), 548-562.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. New York: Springer-Verlag.

Vapnik, V. (1998). *Statistical Learning Theory*, New York: John Wiley.

Yao, X., Crook, J & Andreeva, G. (2015). Support vector regression for loss given default modelling. *European Journal of Operational Research* 240(2), 528-238.

Zhang, J & Thomas, L. (2010). Comparisons of linear regression and survival analysis using single and mixture distributions approaches in modelling LGD. *International Journal of Forecasting* 28(1), 204-215