Journal of Credit Risk **18**(3), 77–103
DOI: 10.21314/JCR.2022.004

Research Paper

Generalized additive modeling of the credit risk of Korean personal bank loans

Young-Ah Kim,[1] Peter G. Moffatt[2] and Simon A. Peters[3]

[1] Essex Business School, 10 Elmer Approach, Southend-on-Sea SS1 1LW, UK;
email: yakim@essex.ac.uk

[2] School of Economics, University of East Anglia, Norwich Research Park,
Norwich NR4 7TJ, UK; email: p.moffatt@uea.ac.uk

[3] School of Social Sciences, University of Manchester, Oxford Road,
Manchester M13 9PL, UK; email: simon.peters@manchester.ac.uk

(Received October 6, 2019; revised October 6, 2021; accepted November 8, 2021)

ABSTRACT

We analyze consumer defaults in a sample of 64 000 customers taking personal loans from a Korean bank. Applying a generalized additive modeling (GAM) framework, we show a nonlinear impact of loan and borrower characteristics. In particular, the likelihood of default is high for both low-income borrowers and high-income borrowers. Our results are robust to a range of different tests, and they highlight the usefulness of the GAM framework, especially the graphical presentation of nonlinearities.

Keywords: generalized additive models; basis splines (B-splines); credit scoring; loan defaults; signal detection theory; misclassification costs.

1 INTRODUCTION

Models for binary outcomes, such as logit and probit, are popular tools in the modeling of loan defaults (see, for example, Greene 1998; Thomas *et al* 2017).

Corresponding author: P. G. Moffatt

Print ISSN 1744-6619 | Online ISSN 1755-9723

Such models are usually classified under the heading of generalized linear models (GLMs). In this paper we consider an extension to the GLM framework: generalized additive models (GAMs) (Hastie and Tibshirani 1990). The essence of the GAM approach is that each individual effect is modeled separately using a nonparametric smoothing procedure, allowing nonlinear effects of each independent variable to be captured. The major attraction of the GAM approach is flexibility: while some parametric approaches require the nature of the nonlinear effects to be specified before estimation, the GAM approach finds the pattern through estimation, and it is therefore not constrained by prior beliefs. Clearly, the greater flexibility of GAMs has the potential to bring about improvements in the ability to discriminate between defaulters and nondefaulters, greater accuracy in predicted default probabilities and lower misclassification costs. An assessment of these potential benefits is the central objective of this paper.

It must be acknowledged that there are many other ways of extending the GLM framework to allow predictor variables to have fully flexible effects on the outcome. Methods that have become popular in recent years include neural networks, decision trees and support vector machines. A survey of such methods is provided by Lessmann *et al* (2015), and recent applications of these methods to credit default data include Khandani *et al* (2010), Butaru *et al* (2016) and Abdou *et al* (2019). A potential problem with all of these approaches (including GAMs) is that of overfitting, and for this reason there is a clear need to pay close attention to out-of-sample predictive performance when evaluating and comparing models.

Larsen (2015) provides a number of compelling reasons for choosing the GAM approach over other approaches that compete in terms of flexibility. One obvious reason is convenience: estimation and testing are both straightforward and transparent in the context of GAMs, leading to a reliable and unambiguous strategy for model selection. A more important reason is interpretability: as will be demonstrated in this paper, the structure of a GAM is such that it is possible to provide a graphical representation of the effect of a single predictive variable on the outcome, and to interpret this effect in ways that appeal to nonspecialists. A further reason is regularization: the chosen smoothing procedure typically contains a smoothing parameter (in our case this will be the number of knots in the spline), which is set in advance and can be used directly to tackle the bias–efficiency trade-off.

Application of the GAM approach to the modeling of loan defaults was originally suggested by Taylan *et al* (2007). More recently, the GAM approach has been applied to default data on small and medium-sized enterprises in Italy by Calabrese and Osmetti (2015) and to default data on retail customers in Germany and on German companies by Lohmann and Ohlinger (2018a,b).

The objective of this paper is to showcase the GAM methodology by applying it to default data from a sample of around 64 000 customers taking out personal loans from a Korean bank, with a variety of loan purposes. To our knowledge, this is the first application of the GAM approach to default data on private borrowers of a commercial bank. The factors for which we find nonlinear effects are characteristics of individual borrowers such as age, income and amount borrowed, and we will focus on these nonlinear effects in the interpretation of the model results. We will highlight the graphical presentations of these nonlinear effects as the most attractive, and useful, feature of the GAM framework. An important feature of our data is that it is "unbalanced" in the sense that only 1.5% of the cases in the estimation sample are observed as defaults. As we shall see, this feature of the data has important implications in the process of assessing model performance.

The smoothing procedure we adopt in the implementation of the GAM approach is the basis spline (B-spline) smoother (de Boor 2001). The B-spline procedure offers an attractive compromise between polynomial regression and kernel regression. The former provides a global fit, in the sense that the position of the smoother at any point is determined by all observations, even those furthest from the point; the latter provides a local fit, in the sense that only local observations determine the position of the smoother at any point. The B-spline smoother lies somewhere in between.

The B-spline approach has another major advantage that is not widely discussed. It is a nonparametric technique that can be performed as a (generalized) linear regression, since it simply amounts to a regression of the dependent variable on a set of basis functions. This clearly makes implementation relatively straightforward. A further advantage is that by-products of regression analysis such as statistical significance tests may be exploited to the full. Statistical testing is often an awkward problem in the context of nonparametric regression or machine learning models, involving nonstandard distributions and/or resampling methods (see, for example, Gu *et al* 2007). Using the B-spline, it becomes possible, using standard regression-based tests, to adjudicate between models, and in particular to make valid judgments on whether a predictor may be represented flexibly at all, in preference to assuming a linear effect.

We estimate a number of models, both GLMs and GAMs, with varying levels of flexibility. We also use a variety of approaches to evaluate the predictive performance of the estimated models. A very useful survey of these techniques is provided by Lessmann *et al* (2015). We start by applying the well-established receiver operating characteristic (ROC) graphical technique. We also supplement the ROC results with the application of precision–recall curve (PRC) techniques. In doing so, we are following the recommendation of Saito and Rehmsmeier (2015) and others, who demonstrate that the PRC may be more informative than the ROC in the presence of unbalanced samples. We then progress to methods that measure the calibration of models (that is, the closeness of predicted probabilities to outcomes). Finally, we consider methods for determining the optimal threshold for rejecting loan applications given information on misclassification costs, and we use total misclassification costs as a further model selection criterion.

In Section 2 we motivate and outline the GAM framework, and describe the data. In Section 3 we present and interpret the results of applying the GAM framework to the data set, and we also apply a range of model evaluation techniques in order to compare the performance of GAMs to that of less flexible models. Section 4 provides a summary of the findings.

2 MODELING STRATEGY AND DATA

2.1 GAMs

Traditional regression models frequently fail for the simple reason that the effects of interest are often nonlinear. To characterize such effects, flexible statistical methods such as nonparametric regression are a useful first step (Fox 2002). However, if the number of independent variables is large, many forms of nonparametric regression do not perform well. Moreover, in a framework of nonparametric regression, it is more difficult to interpret results, to perform significance tests and to make predictions. To overcome these difficulties, Stone (1985) proposed using additive models. These models estimate an additive approximation of the multivariate regression function. For noncontinuous (eg, binary) outcomes, further generality is required. Hastie and Tibshirani (1990) introduced the framework of GAMs. These models include a link function that allows for the discrete nature of the dependent variable.

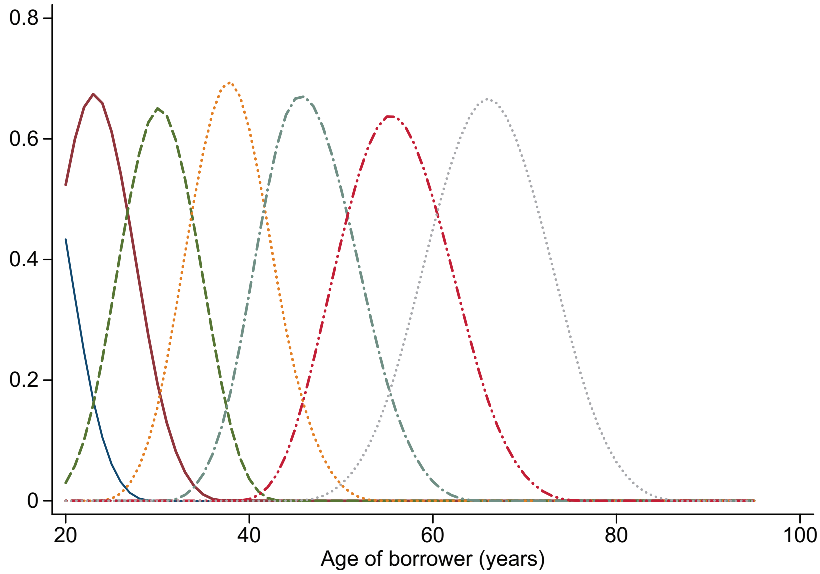
For the case of a binary dependent variable y_i and a total of m available predictors, if we assume a logit link function, the model takes the following form:

$$P(y_i = 1 \mid x_{1i}, \dots, x_{mi}) = \Lambda\left(\beta_0 + \sum_{j=1}^{m} f_j(x_{ji})\right), \tag{2.1}$$

where $\Lambda(\cdot)$ is the logistic function $\Lambda(u) = \exp(u)/(1 + \exp(u))$. In general the functions f_j are piecewise polynomials (or “splines”), although they do not have to be. Some predictors are better modeled linearly (as with $f_j(x_{ji}) = \beta_j x_{ji}$ in the context of (2.1)) and, of course, some can only be modeled in this way (eg, binary dummy variables).

Splines form a useful compromise between the global fit of polynomial regression and the local fit of kernel smoothers. The “pieces” of the piecewise polynomials are separated by a sequence of K “knots”, $\chi_1, \dots, \chi_K$, and they are forced to join smoothly at these knots. Cubic splines are usually chosen, and the smoothness requirement is that the piecewise cubic functions are continuous and have continuous

FIGURE 1 Basis functions for the variable “age”; knots at 23, 30, 38, 44, 55.



first and second derivatives at the knots. This will guarantee that the spline appears smooth when viewed, since, according to Hastie and Tibshirani (1990, p. 22), “our eyes are skilled at picking up second and lower order discontinuities, but not higher”.

The more knots used, the more flexible the smoother. However more knots also means more parameters to estimate, and therefore fewer degrees of freedom. Clearly, the choice of the number of knots must depend on the sample size: the larger the sample, the more knots that can be used. Another choice that needs to be made is the positioning of the knots. The approach adopted here is to place knots at appropriate quantiles of the predictor variable.

The most popular approach for obtaining a piecewise cubic smoother with the required properties is the B-spline approach (de Boor 2001). This amounts to a linear regression of the dependent variable on a set of basis functions. If there are K knots, there are $K + 4$ basis functions in total, although for practical reasons only $K + 2$ of them are used in the regression. For illustration, Figure 1 shows the basis functions obtained from the variable “age”, which is used in the models of later sections. There are five knots, and therefore seven basis functions are used in the regression. The basis functions may be computed using a method developed by Newson (2000).

If the basis functions to be used in the B-spline regression are $B_1(\cdot), \dots, B_{K+2}(\cdot)$, then the piecewise cubic functions $f_j(\cdot)$ appearing in (2.1) may be expressed as

$$f_j(x_{ji}) = \sum_{k=1}^{K+2} \gamma_{jk} B_k(x_{ji}), \quad j = 1, \dots, m. \tag{2.2}$$

It is important that (2.2) does not contain an intercept. This is necessary for the model intercept (β_0 in (2.1)) to be identified.

Note also that (2.2) would lead to a fully general GAM in the sense that all m of the predictors are assumed to have flexible effects. As noted in the discussion following (2.1), there are strong reasons for not modeling the effect of every predictor in accordance with (2.2).

As mentioned in Section 1, a little-recognized advantage to using GAMs in combination with B-splines is that the approach allows for estimation via a linear regression. Although the regression coefficients on the basis functions (ie, the estimates of the γ_{jk} in (2.2)) are themselves hard to interpret, it is straightforward to perform regression-based tests (eg, for predictor j, a joint test of $\mathrm{H}_0: \gamma_{j1} = 0, \dots, \gamma_{j,K+2} = 0$) in order to, first, test for the presence of nonlinear effects of predictors and, second, adjudicate between models.

2.2 Data

The data is from a Korea-based commercial bank, with branches distributed throughout South Korea. The bank is engaged in the provision of a wide range of commercial and consumer banking services. This research focuses on individual loans, and each unit of observation is an individual borrower. There are a total of 64 579 borrowers in the data set. The sample size is therefore considerably higher than the average sample size of the 48 similar studies considered by Lessmann *et al* (2015), which was 6167.

Of the 64 579 borrowers, 32 534 fail to report income. Since income is one of the key determinants of default, the estimation sample consists of only the 32 045 borrowers for whom income is known. Although this may appear to be a large estimation sample, the unbalanced nature of the sample limits the efficiency of estimation, and this is why we choose to use all observations for which income is observed. The remainder of the sample provides a convenient test sample. Out-of-sample predictive methods will be performed on this test sample. The measure of income used for the test sample is imputed using the method of Afifi and Elashoff (1966): for the estimation sample, a linear regression of income is performed on all other independent variables, and predictions from this regression are applied to the test sample. Note that using a test sample with missing data somewhat complicates the task of out-of-sample prediction, but this applies to for all models, so our model-comparison procedure remains valid. In any case, the problem of predicting default with incomplete information on borrowers is surely a problem with real-world relevance.

We present descriptive statistics for all variables used in the analysis in the online appendix (see Tables A1 and A2 and the variable definitions provided in Table A3). Figure A1 in the online appendix presents histograms showing the distributions of the three key continuous independent variables used in the models of the next section: amount borrowed, income and age. The first two of these show a strong positive skew, leading us to apply the logarithmic transformation before estimation.

The "purpose of loan" data is divided into five categories: "unspecified", "property", "financial", "living" (ie, living expenses) and "other". Definitions of these categories, in terms of the subcategories they contain, are provided in Table A3 (online). Table A4 (online) shows descriptive statistics for loan amounts separately for each category. There, we see that property loans tend to be the largest, while loans for living expenses tend to be the smallest. However, we also see that the property loans account for only a small proportion of the total loans, and the majority of loans in the sample are unsecured.

All loans commenced between May 1992 and June 2012, with dates of redemption between June 2001 and February 2051. The reference date for the default information is December 31, 2012; "default" is defined as any sort of failure to meet the obligations of the loan between commencement of the loan and this date. Clearly, the length of time over which the loan is observed is an important determinant of the probability of default being observed, and this variable ("duration", measured in years) is always included in our estimation.

Operationally, banks in Korea aim to achieve an overall default rate for individual loans of between 1.5% and 2.0%.[1] As can be seen from the first row of Table A1 (online), the overall default rate in our estimation sample is marginally below 1.5%. However, note from Table A2 (online) that the default rate in the test sample is a much higher 4.3%. Recalling that the test sample consists of borrowers for whom income is not observed, this difference clearly suggests that missing income is a strong risk factor, and it also presents a challenge for out-of-sample prediction.

[1] See https://bit.ly/3zTykFd.

3 RESULTS

3.1 Model estimates

The dependent variable in our analysis is 1 if the borrower defaulted and 0 otherwise. The predictors are "log(amount)";[2] "log(income)"; "age" (in years); "male" (1 if male, 0 if female); "married" (1 if married, 0 otherwise); "dependants" (the number of dependents); and four dummy variables for the purpose of the loan: "property", "financial", "living", "other" and (the base case) "unspecified". There is also a variable, "duration", representing the duration of the loan, measured in years.[3] The estimation sample of 32 045 borrowers is used for the estimation of all models.

A useful model-free framework in which to uncover preliminary evidence of nonlinear effects is weight of evidence (WoE) coding (see Larsen 2016). One important aspect in which WoE scores are model-free is their invariance to monotonic transformations of the independent variable; hence, WoE scores reveal the same nonlinearities whether, for example, income or log(income) is being considered. WoE scores for categorized versions of age, log(amount) and log(income), obtained after splitting each into ten contiguous categories, are presented in Figure 2. The scores, and plots, were obtained using the R package `Information` (Larsen 2016). None of the three WoE score vectors show a monotonic change as the category values increase. This nonmonotonic behavior is usually interpreted as an indication that these variables may affect the default probability in a nonlinear fashion. A further output of WoE analysis is the information value (IV), which acts as a summary of the WoE over all categories. The IVs of all variables are shown in the final column of Table A1 in the online appendix. According to a rule of thumb,[4]

- an IV of less than 0.02 indicates a useless predictor, one less than 0.10 indicates a weak predictor;
- an IV of between 0.10 and 0.3 indicates a "medium" predictor; and
- an IV of greater than 0.3 indicates a strong predictor.

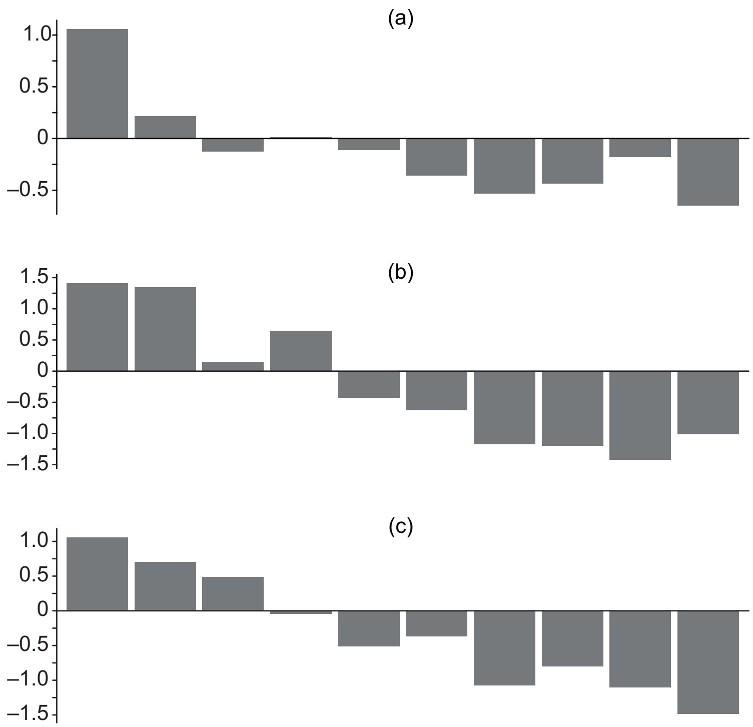
By this rule of thumb, we see that both log(amount) and log(income) are strong predictors, while age is a medium predictor.

[2] We acknowledge that there may be an element of endogeneity in the variable "amount", since the size of the approved loan will likely be determined partly by the perceived creditworthiness of the borrower. The broadly negative effect of this variable seen in both Figure 2 and Figure 3 is consistent with this possibility.

[3] See Table A3 in the online appendix for further disaggregation of these purpose categories.

[4] URL: www.listendata.com/2015/03/weight-of-evidence-woe-and-information.html.

FIGURE 2 WoE score vectors for (a) age, (b) log(amount) and (c) log(income).



Six binary logit models have been estimated, and the results are reported in Table 1. Model 1 is a simple logit model with linear effects only. Model 2 is a simple logit model with quadratic effects assumed for the three continuous predictors (log(amount), log(income) and age). On the basis of the Akaike information criterion (AIC), model 2 is superior to model 1. On the basis of t-statistics, while it is clear in model 2 that both age and log(income) have quadratic effects on the default probability, there is no evidence of a quadratic effect of log(amount).

Models 3–6 are GAMs. Models 3–5 assume flexible effects for log(amount), log(income) and age, respectively. On the basis of the AIC, these models are all superior to models 1 and 2, confirming that flexible specifications for these variables are desirable. It is particularly interesting that the strong nonlinear effect of log(amount) seen in model 3 was not picked up by the quadratic specification in model 2. Finally, model 6 assumes flexible effects for all three variables (log(amount), log(income) and age). Once again using the AIC, we can see that this model is the best performer

TABLE 1 Results of six binary logit models of loan default. [Table continues on next three pages.]

	Model 1 (baseline)	**Model 2 (quadratic)**	**Model 3 (flex-amount)**	**Model 4 (flex-income)**	**Model 5 (flex-age)**	**Model 6 (flex-all)**
Log(amount)	−0.363***	−0.802	—	−0.811	−0.847	—
	(−5.96)	(−1.54)	—	(−1.59)	(−1.63)	—
Log(amount)^2	—	0.0158	—	0.0167	0.0171	—
	—	(0.95)	—	(1.02)	(1.03)	—
Log(income)	−0.188***	0.909**	0.785*	—	0.969**	—
	(−5.82)	(2.79)	(2.41)	—	(2.95)	—
Log(income)^2	—	−0.0394***	−0.0345**	—	−0.0414***	—
	—	(−3.38)	(−2.95)	—	(−3.53)	—
Age	−0.0266***	−0.147***	−0.140***	−0.138***	—	—
	(−4.07)	(−3.90)	(−3.70)	(−3.66)	—	—
Age^2	—	0.00139***	0.00133**	0.00130**	—	—
	—	(3.33)	(3.15)	(3.10)	—	—
Male	0.449***	0.517***	0.520***	0.563***	0.515***	0.559***
	(4.23)	(4.80)	(4.82)	(5.19)	(4.76)	(5.11)
Married	−0.120	0.0153	0.0251	0.0258	0.0411	0.0577
	(−0.95)	(0.12)	(0.19)	(0.20)	(0.30)	(0.42)
Dependants	−0.0402	−0.0324	−0.0335	−0.0291	−0.0351	−0.0323
	(−1.44)	(−1.16)	(−1.20)	(−1.05)	(−1.26)	(−1.16)
Property	−0.0646	−0.171	−0.112	−0.204	−0.151	−0.0938
	(−0.17)	(−0.45)	(−0.28)	(−0.53)	(−0.39)	(−0.23)
Financial	2.484***	2.437***	2.360***	2.387***	2.439***	2.314***
	(12.54)	(12.20)	(11.59)	(11.92)	(12.18)	(11.31)

TABLE 1 Continued.

	Model 1 (baseline)	**Model 2 (quadratic)**	**Model 3 (flex-amount)**	**Model 4 (flex-income)**	**Model 5 (flex-age)**	**Model 6 (flex-all)**
Living	1.872***	1.833***	1.697***	1.791***	1.813***	1.644***
	(16.46)	(15.90)	(14.41)	(15.52)	(15.65)	(13.88)
Other	0.104	0.146	0.284	0.195	0.156	0.336
	(0.26)	(0.34)	(0.71)	(0.46)	(0.37)	(0.84)
Duration	0.102**	0.112**	0.104**	0.116**	0.115**	0.111**
	(2.76)	(3.01)	(2.80)	(3.12)	(3.09)	(2.96)
$\gamma_{\text{log(amount)},1}$	—	—	0.186	—	—	0.187
	—	—	(0.32)	—	—	(0.33)
$\gamma_{\text{log(amount)},2}$	—	—	0.445	—	—	0.422
	—	—	(1.04)	—	—	(0.98)
$\gamma_{\text{log(amount)},3}$	—	—	−0.733	—	—	−0.707
	—	—	(−1.62)	—	—	(−1.56)
$\gamma_{\text{log(amount)},4}$	—	—	−0.367	—	—	−0.340
	—	—	(−0.84)	—	—	(−0.77)
$\gamma_{\text{log(amount)},5}$	—	—	−1.258*	—	—	−1.180*
	—	—	(−2.28)	—	—	(−2.13)
$\gamma_{\text{log(amount)},6}$	—	—	−1.527*	—	—	−1.535*
	—	—	(−2.05)	—	—	(−2.05)
$\gamma_{\text{log(amount)},7}$	—	—	−0.369	—	—	−0.338
	—	—	(−0.37)	—	—	(−0.34)

TABLE 1 Continued.

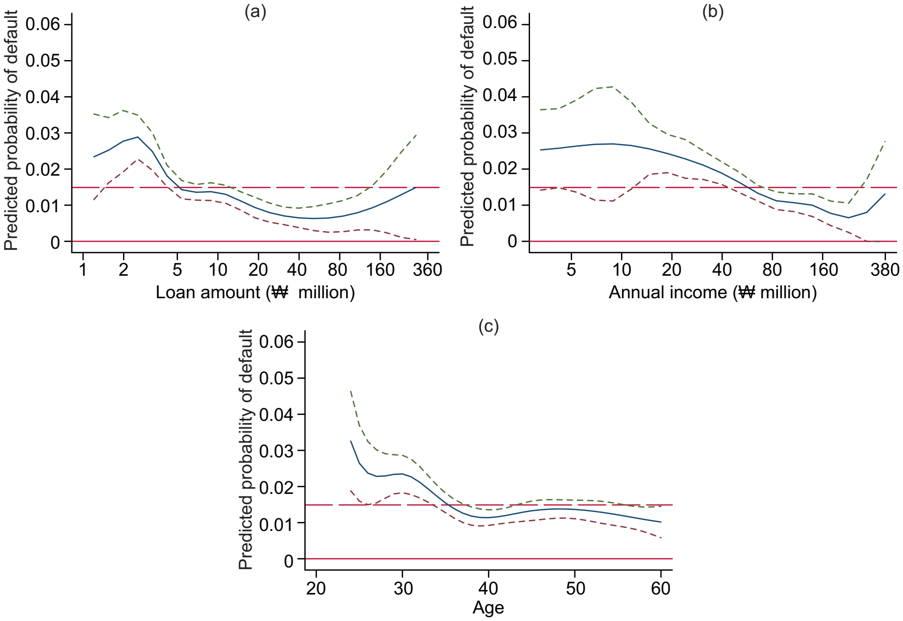
	Model 1 (baseline)	**Model 2 (quadratic)**	**Model 3 (flex-amount)**	**Model 4 (flex-income)**	**Model 5 (flex-age)**	**Model 6 (flex-all)**
$\gamma_{\log(\text{income}),1}$	—	—	—	0.0917	—	0.139
	—	—	—	(0.14)	—	(0.21)
$\gamma_{\log(\text{income}),2}$	—	—	—	−0.0782	—	−0.0442
	—	—	—	(−0.19)	—	(−0.11)
$\gamma_{\log(\text{income}),3}$	—	—	—	−0.331	—	−0.318
	—	—	—	(−0.91)	—	(−0.87)
$\gamma_{\log(\text{income}),4}$	—	—	—	−1.106**	—	−0.983**
	—	—	—	(−3.19)	—	(−2.80)
$\gamma_{\log(\text{income}),5}$	—	—	—	−0.879*	—	−0.811
	—	—	—	(−1.97)	—	(−1.81)
$\gamma_{\log(\text{income}),6}$	—	—	—	−1.670**	—	−1.599*
	—	—	—	(−2.68)	—	(−2.53)
$\gamma_{\log(\text{income}),7}$	—	—	—	−1.336	—	−1.123
	—	—	—	(−1.16)	—	(−0.97)
$\gamma_{\text{age},1}$	—	—	—	—	7.189*	7.288*
	—	—	—	—	(2.45)	(2.45)
$\gamma_{\text{age},2}$	—	—	—	—	−0.120	−0.184
	—	—	—	—	(−0.08)	(−0.12)
$\gamma_{\text{age},3}$	—	—	—	—	0.963	0.953
	—	—	—	—	(0.65)	(0.63)

TABLE 1 Continued.

	Model 1 (baseline)	Model 2 (quadratic)	Model 3 (flex-amount)	Model 4 (flex-income)	Model 5 (flex-age)	Model 6 (flex-all)
$\gamma_{\text{age},4}$	—	—	—	—	−0.546	−0.502
	—	—	—	—	(−0.38)	(−0.34)
$\gamma_{\text{age},5}$	—	—	—	—	−0.00999	0.0297
	—	—	—	—	(−0.01)	(0.02)
$\gamma_{\text{age},6}$	—	—	—	—	−0.200	−0.133
	—	—	—	—	(−0.15)	(−0.10)
$\gamma_{\text{age},7}$	—	—	—	—	−0.673	−0.706
	—	—	—	—	(−0.35)	(−0.36)
Constant	5.100***	3.103	−4.560	7.533	−0.484	−4.088**
	(4.79)	(0.65)	(−1.86)	(1.83)	(−0.10)	(−2.62)
LogL	−2135.2	−2123.3	−2107.2	−2117.1	−2117.5	−2096.5
n	32 045	32 045	32 045	32 045	32 045	32 045
k	12	15	20	20	20	30
AIC	4294.5	4276.6	4254.4	4274.2	4275.0	4253.0

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$. Values in parentheses are t-statistics. Model 1: linear effects only. Model 2: quadratic effects. Model 3: a GAM with a flexible effect of log(amount) (knots: 14.9, 15.4, 16.1, 16.8, 18.1). Model 4: a GAM with a flexible effect of log(income) (knots: 15.9, 16.8, 17.3, 17.8, 18.2). Model 5: a GAM with a flexible effect of age (knots: 23, 30, 38, 44, 55). Model 6: a GAM with flexible effects of all three. LogL, maximized loglikelihood. AIC $= 2(k - \text{LogL})$ (where k is the number of parameters) is a measure of model fit. The best-fitting model is the one with the lowest AIC. Parameters $\gamma_{j,k}$, $j = \{\log(\text{amount}), \log(\text{income}), \text{age}\}$, $k = 1, \dots, 7$, are the coefficients on the basis functions used to obtain the B-spline (see (2.2)).

FIGURE 3 Predicted probability of default against (a) amount, (b) income and (c) age.



Other variables are set to means. Loan amounts and income are given on a logarithmic scale. Solid curves represent predicted probabilities; short-dashed curves represent 95% confidence bands; long-dashed lines are drawn at the sample proportion of defaults.

of the six models, and its superiority over models 3–5 vindicates the assumption of a flexible effect for all three variables.

It is well known that the coefficients on the basis functions are hard to interpret. However, it is relatively straightforward to use the estimated coefficients to generate curves, with confidence bands, for predicted probabilities of default against each of the variables for which flexible effects are assumed. To obtain the confidence bands, a loop is performed over a suitable range of values of the x-variable of interest. At each stage of the loop, a prediction interval is obtained using the basis functions evaluated at the current value of the x-variable, with all other predictors set to their means. The resulting plots are shown in Figure 3. Note that log scales have been used in parts (a) and (b), to facilitate interpretation. These plots clearly confirm the highly nonlinear nature of the effects detected in the WoE analysis reported at the start of this section.

In the cases of log(amount) and log(income), the most striking features of the plots are the pronounced upticks at the upper end of the scale. Possible explanations for these upticks are given in the economics literature. For example, loans offered to borrowers with the highest reported incomes may be "liar's loans" (Jiang *et al* 2014): borrowers who falsely report income tend to report higher levels of income, and, of course, any sort of falsification of borrower characteristics must be associated with higher credit risk. This would provide an explanation for the uptick in the case of income. Regarding the case of loan amount, we may appeal to the literature on "strategic defaulters": these are borrowers who choose to default if they perceive the net benefit of default to be positive. It is well known from this literature (see, for example, Bradley *et al* 2015) that strategic default is strongly associated with larger loan amounts, thus providing an explanation for the uptick.

In the case of age, we see an overall downward trend in default probability; it appears that older Korean borrowers are less likely to default. However, Figure 3(c) also shows hints that the effect of age takes the form of a downward step function, leveling off at particular stages of the life cycle, namely, the late twenties and early forties. These age ranges correspond to the definitive stages in the life cycle identified in the Korean context by Kim and Lee (2010). We suggest that researchers look out for this sort of pattern in future research.

If we now turn to the effects of the other predictors, we can see that (ceteris paribus) male borrowers are significantly more likely to default, while marital status and the number of dependents are apparently unimportant. The purpose of the loan is important, with those borrowing for either financial purposes (which include the subcategories "business", "investment", "repayment of other loans" and "repayment of credit card") or living expenses being significantly more likely to default than those borrowing for unspecified purposes, property-related purposes or other purposes. Finally, we see that the duration of the loan has the expected positive effect on the probability of default.

A well-known limitation to this sort of analysis that should be acknowledged is that of sample selection bias (see Greene 1998). Clearly, the data set consists of the loan applications that have been approved by the bank, and hence the estimation results must be sensitive to the credit scoring algorithm that the bank is using. The results reported above should therefore be interpreted conditionally. Of course, in the present setting there is nothing that can be done to address this problem, first because the bank's credit scoring algorithm is unknown to us, and second because the available data set contains no information on loans that were declined.

3.2 Predictive performance

In Section 3.1 a number of models were estimated, and they were then compared using the AIC. This led to the conclusion that the most flexible GAM model (model 6) was the most preferred model. There are many other ways of assessing the predictive performance of binary-data models. In this section we will apply a range of these techniques.

The various techniques that we consider each fall under one of three headings: discrimination, calibration and misclassification-cost minimization. We apply each technique both in-sample and out-of-sample. For the in-sample predictions, we use the estimation sample, consisting of 32 045 observations, for both estimation and prediction. For the out-of-sample predictions, we use the estimation sample for estimation, and we use the test sample, consisting of 32 534 observations (with missing income imputed), for prediction.

3.2.1 Discrimination

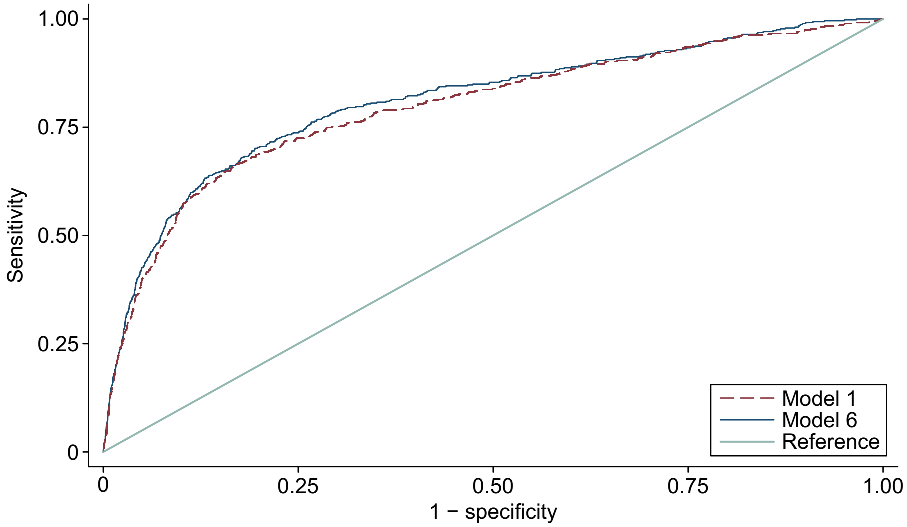
"Discrimination" refers to the ability of a model to separate defaulters from nondefaulters. Methods of assessing discrimination sometimes come under the heading of signal detection theory. These methods are very popular in areas such as medicine (Saito and Rehmsmeier 2015) and criminology (Mumpower and McClelland 2014). The methods have been applied in credit scoring applications by Adams and Hand (1999), Medema *et al* (2009), Yu *et al* (2009), Hand and Anagnostopoulos (2013), Castermans *et al* (2010) and Lohmann and Ohlinger (2018a).

The most widely used of these techniques is the ROC curve. To assess the discriminatory performance of a model, the predicted probability of default for each borrower is obtained, and it is taken that default is predicted whenever the predicted probability is above a threshold. The ROC curve is a plot of the true positive rate (TPR, or "sensitivity") against the false positive rate (FPR, or 1 minus "specificity") for all possible values of the threshold. For a model that is useless in discrimination, the ROC curve will lie close to the 45° line. The higher above the 45° line the ROC curve lies, the better the discriminatory performance of the model, since this implies that the TPR is rising faster than the FPR when the threshold is lowered. A model that predicts perfectly (with the predicted probabilities perfectly separating the two groups) will produce an ROC curve consisting of a vertical line and a horizontal line meeting at the top-left corner of the graph.

Given that the height of the ROC curve is an indicator of discriminatory performance, a natural quantitative measure of this performance is the area under the ROC curve ($\mathrm{AUC}_{\mathrm{ROC}}$). According to Hosmer *et al* (2013),

- an $\mathrm{AUC}_{\mathrm{ROC}}$ of 0.5 indicates no discrimination,
- an $\mathrm{AUC}_{\mathrm{ROC}}$ between 0.7 and 0.8 indicates acceptable discrimination,
- an $\mathrm{AUC}_{\mathrm{ROC}}$ between 0.8 and 0.9 indicates excellent discrimination, and
- an $\mathrm{AUC}_{\mathrm{ROC}}$ greater than 0.9 indicates outstanding discrimination.

FIGURE 4 ROC curves from models 1 and 6.



$\mathrm{AUC}_{\mathrm{ROC}} = 0.796$ for model 1 and $\mathrm{AUC}_{\mathrm{ROC}} = 0.809$ for model 6.

As noted by Adams and Hand (1999), $\mathrm{AUC}_{\mathrm{ROC}}$ can intuitively be interpreted as the probability that a randomly chosen default case will have been ranked higher by the model than a randomly chosen nondefault case. There is also a close link between $\mathrm{AUC}_{\mathrm{ROC}}$ and the Gini coefficient, which is widely used in the economics literature as a measure of inequality (see Schechtman and Schechtman 2019).

Figure 4 shows the (in-sample) ROC curves obtained from the simplest model (model 1) and the most flexible model (model 6). The areas under the two curves (the $\mathrm{AUC}_{\mathrm{ROC}}$) are, respectively, 0.796 and 0.809, and in the terminology of Hosmer *et al* (2013) both of these values lie at the cusp between acceptable and excellent discrimination. The difference between the two $\mathrm{AUC}_{\mathrm{ROC}}$ is clearly very slight, and it is important to consider whether this difference is statistically significant. In developing a test with this objective, it is very important to take account of the strong positive dependence between two different $\mathrm{AUC}_{\mathrm{ROC}}$ obtained using the same data set. To address this issue, we follow Robin *et al* (2011) by bootstrapping the difference between the two $\mathrm{AUC}_{\mathrm{ROC}}$ in order to compute the standard deviation of this difference. The test statistic is then obtained as the ratio of the observed difference to this bootstrapped standard deviation.

Table 2 contains the $\mathrm{AUC}_{\mathrm{ROC}}$, and all of the other model selection criteria discussed in Section 3.2, for all six of the models estimated in Section 3.1, applied both

TABLE 2 Model selection criteria for the six models estimated in Section 3.1. [Table continues on next page.]

(a) Model selection criteria

	Model 1 (baseline)	**Model 2 (quadratic)**	**Model 3 (flex-amount)**	**Model 4 (flex-income)**	**Model 5 (flex-age)**	**Model 6 (flex-all)**
LogL	−2135.2†	−2123.3	−2107.2	−2117.1	−2117.5	−2096.5*
AIC	4294.5†	4276.6	4254.4	4274.2	4275.0	4253.0*

(b) In-sample

	Model 1 (baseline)	**Model 2 (quadratic)**	**Model 3 (flex-amount)**	**Model 4 (flex-income)**	**Model 5 (flex-age)**	**Model 6 (flex-all)**
AUC_{ROC}	0.796†	0.805	0.805	0.807	0.805	0.809*
p-value	0.003	0.186	0.209	0.184	0.160	—
AUC_{PRC}	0.088†	0.091	0.097	0.089	0.093	0.097*
p-value	0.057	0.114	0.900	0.027	0.183	—
MSE	0.01411†	0.01409	0.01401*	0.01410	0.01408	0.01403
R^2_{LE}	0.04118†	0.04306	0.04827*	0.04191	0.04345	0.04697
Cost(10)	4353.1	4285.7	4217.0*	4404.2†	4384.0	4238.4
Cost(15)	5919.6	5974.3†	5780.4	5924.6	5952.6	5707.8*
Cost(20)	7401.7†	7216.1	7143.9	7236.6	7130.6	7105.4*

TABLE 2 Continued.

(c) Out-of-sample

	Model 1 **(baseline)**	**Model 2** **(quadratic)**	**Model 3** **(flex-amount)**	**Model 4** **(flex-income)**	**Model 5** **(flex-age)**	**Model 6** **(flex-all)**
$\mathrm{AUC}_{\mathrm{ROC}}$	0.780*	0.772†	0.774	0.772	0.775	0.776
p-value	0.940	0.019	0.100	0.010	0.336	—
$\mathrm{AUC}_{\mathrm{PRC}}$	0.130	0.129†	0.132	0.132	0.132	0.138*
p-value	0.038	0.011	0.104	0.041	0.082	—
MSE	0.03949	0.03956†	0.03951	0.03939	0.03955	0.03936*
R^2_{LE}	0.04307	0.04128†	0.04247	0.04528	0.04140	0.04611*
Cost(10)	11 374.4	11 372.5	11 362.1	11 311.0*	11 474.1†	11 363.2
Cost(15)	14 275.4†	14 281.8	14 182.2	14 251.5	14 220.8	14 167.8*
Cost(20)	16 566	16 612†	16 556	16 599	16 365	16 230*

(a) Key model selection criteria as reported in Table 1. (b) Measures of in-sample predictive performance (the estimation sample of 32 045 is used for both estimation and prediction). (c) Measures of out-of-sample predictive performance (the estimation sample of 32 045 observations is used for estimation; the test sample of 32 534 is used for prediction). $\mathrm{AUC}_{\mathrm{ROC}}$, area under the ROC curve; $\mathrm{AUC}_{\mathrm{PRC}}$, area under the PRC. The p-values appearing in the row below the $\mathrm{AUC}_{\mathrm{ROC}}$ (respectively, $\mathrm{AUC}_{\mathrm{PRC}}$) are for the (one-tailed) test of the difference between the model's $\mathrm{AUC}_{\mathrm{ROC}}$ (respectively, $\mathrm{AUC}_{\mathrm{PRC}}$) and that of model 6 (using 201 bootstrap replications). MSE, mean squared error (also known as the Brier score). R^2_{LE}, Lave–Effron R-squared measure. Cost($\cdot$), total misclassification cost when the number in parentheses is the cost ratio and the cost of a false positive is 1. *Best-performing model for each criterion. †Worst-performing model for each criterion.

in-sample and out-of-sample. The row below the AUC_{ROC} contains the p-values for (one-sided) tests of the equality of each AUC_{ROC} with that of model 6, implemented using the procedure outlined in the previous paragraph, with a low p-value indicating the superiority of model 6. Model 6 has the highest in-sample AUC_{ROC}. On the basis of the p-values, it is significantly higher than that of model 1, although the differences to other models are not significant.

Out-of-sample, we see that, surprisingly, model 1 has a higher AUC_{ROC} than model 6, although the difference is not significant. Model 6 has a higher AUC_{ROC} than all other models, and the difference is significant in some cases. The apparently superior out-of-sample discriminatory performance of model 1 could suggest overfitting by the other models, but similar indications are not seen anywhere else in the results.

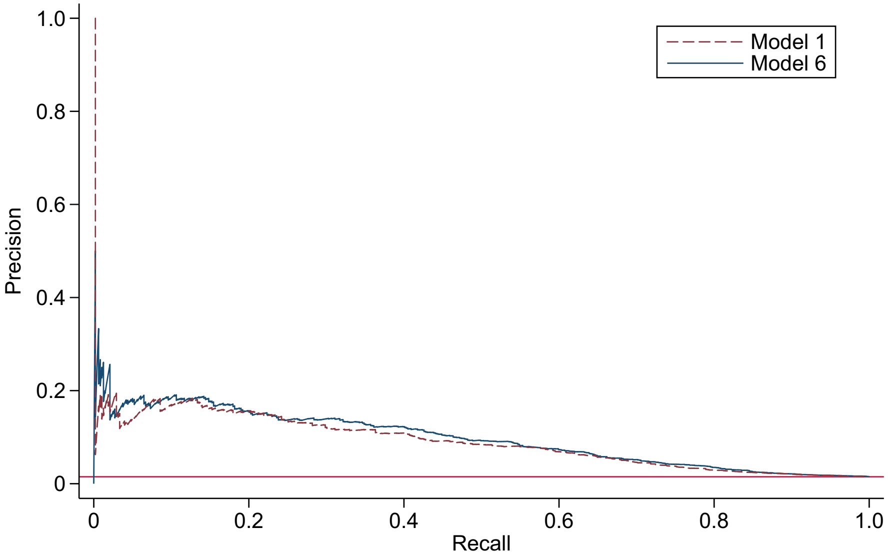
It has been argued by Saito and Rehmsmeier (2015, 2017) that the ROC approach used above may give misleading results when applied to an unbalanced data set. According to Saito and Rehmsmeier (2015, p. 1), "visual interpretability of ROC plots in the context of imbalanced data sets can be deceptive with respect to conclusions about the reliability of classification performance, owing to an intuitive but wrong interpretation of specificity". This is highly relevant to our data set, since only 1.5% of our estimation sample are defaults. Saito and Rehmsmeier suggest that the PRC be used in addition to the ROC curve in the presence of unbalanced data. The PRC is a plot of the positive predicted value (PPV) against the TPR, where the PPV is defined as the number of true positives as a proportion of the total number of positives. In other words, the PPV is the proportion of predicted defaulters (based on model observations) who are true defaulters, while the TPR is the proportion of true defaulters who are correctly predicted to be defaulters. The PRC therefore represents the trade-off between these two proportions. Again, the area under this curve, the AUC_{PRC}, will provide a numerical measure of discriminatory performance. Both the PRC and the AUC_{PRC} can be obtained using a procedure developed by Cook and Ramadas (2020).

Figure 5 shows the (in-sample) PRCs for models 1 and 6. The AUC_{PRC} for all models are presented in Table 2, along with p-values showing comparisons with model 6. These p-values are obtained using a bootstrapping procedure similar to that described above for the comparison of the AUC_{ROC}. Both in-sample and out-of-sample, model 6 has the highest AUC_{PRC}, and in the out-of-sample case some of the differences are significant.

3.2.2 Calibration

The second type of measure of predictive performance comes under the heading of calibration. "Calibration" refers to the closeness with which predicted probabilities

FIGURE 5 PRCs from models 1 and 6.



$\text{AUC}_{\text{PRC}} = 0.088$ for model 1 and $\text{AUC}_{\text{PRC}} = 0.097$ for model 6.

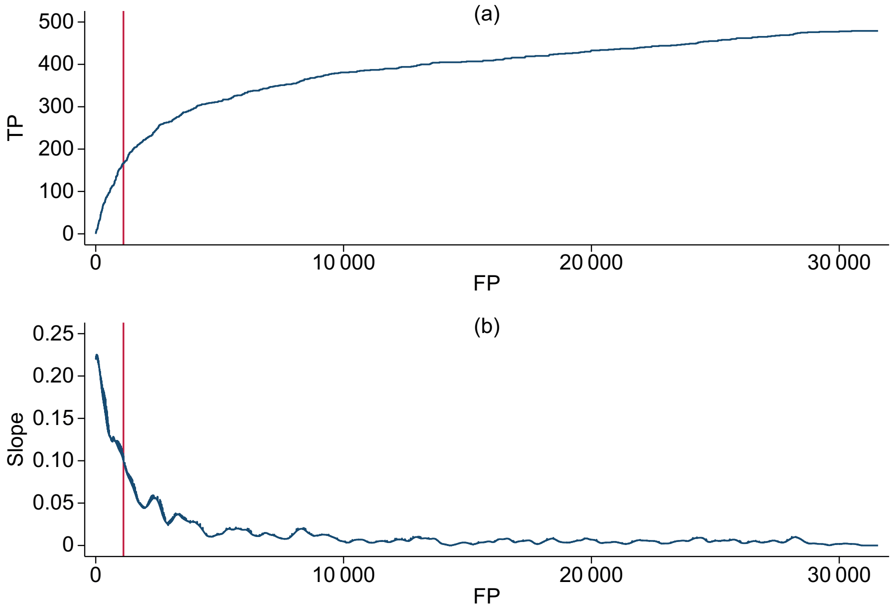
correspond to actual outcomes. Calibration is clearly important if accurate estimation of default probabilities is what is required.

One obvious calibration measure is the mean squared error (MSE) – also known as the Brier score in the classification literature – which is obtained simply as the mean of squared differences between the binary outcome and the predicted probability. Closely related to the MSE is the Lave–Effron R-squared measure, R^2_{LE} (Lave 1970; Efron 1978). For each of the models estimated in Section 3.1, we report both the MSE and the R^2_{LE} in Table 2. In-sample, model 3 (flex-amount) performs best on both of these criteria, while out-of-sample, model 6 (flex-all) performs best on both. This is reassuring: the apparently superior out-of-sample performance of the most general model suggests that overfitting is not an issue.

3.2.3 Misclassification-cost minimization

All of the evaluation measures considered so far are overall measures of the predictive performance of a model and are not specific to a single classification rule. The final question we address in this section is twofold: how to select a classification rule, and how to then assess the performance of the model conditional on the chosen rule.

FIGURE 6 TP/FP ratio and the gradient of the TP/FP ratio for model 6.



(a) Plot of the number of true positives against the number of false positives from model 6. (b) Slope of the graph shown in part (a). The vertical line appears at the number of false positives at which the slope is equal to 0.1.

The key to finding the optimal classification rule is to have information on misclassification costs. Here, we will follow convention by making the following assumptions. The cost of making a correct prediction is zero; that is, if a nondefaulter is correctly classified as a nondefaulter (a true negative), or if a defaulter is correctly classified as a defaulter (a true positive), then the cost is zero. However, the cost of misclassification is positive, and, for a number of reasons, the cost of incorrectly classifying a defaulter as a nondefaulter (a false negative) is assumed to be higher than the cost of incorrectly classifying a nondefaulter as a defaulter (a false positive).

Adams and Hand (1999) emphasize the uncertainty surrounding the misclassification cost ratio, but, on the basis of discussions with banking domain experts, they arrive at a range of 6 to 15, with a most likely value around 10. Abdou *et al* (2019) choose cost ratios in a similar range, again on the basis of guidance from bank officials. Lessmann *et al* (2015) consider a wider range of cost ratios, from 2 to 50. Here, we will assume three different cost ratios: 10, 15 and 20. We do not consider cost ratios lower than 10, because doing so tends to lead to a rule whereby all loans are approved.

TABLE 3 Confusion table for the threshold 0.0758 (optimal for model 6 assuming a cost ratio of 10).

	$p < \mathbf{0.0758}$	$p \geqslant \mathbf{0.0758}$
Nondefault	TN = 30 448	FP = 1118
Default	FN = 312	TP = 167

Total sample size: 32 045.

Adams and Hand (1999) also recommend a method for finding the cost-minimizing cutoff using the ROC.[5] Here, we apply a more direct approach, which we consider to be more intuitive and which makes clear the role of the marginal concept in the solution of the optimization problem. In Figure 6(a) we plot the number of true positives (TP) against the number of false positives (FP) from model 6. This graph is closely related to the ROC graph for model 6 shown in Figure 4, but it should be stressed that here we are using the number of cases instead of the proportion of cases. Consider the interpretation of the slope at any point on the graph shown in Figure 6(a). This slope represents the number of true positives that can be gained in exchange for one false positive, or the marginal benefit from a false positive. Movement up the curve from this point will reduce total costs provided that the saving resulting from the increase in true positives exceeds the increase in costs resulting from the false positive (that is, provided that the marginal benefit of a false positive exceeds the marginal cost). Hence, movement up the curve at any point will reduce costs if the slope of the curve at that point exceeds the reciprocal of the cost ratio.

With the assumption of a cost ratio of 10, we therefore have the following cost-minimization rule: find FP corresponding to a slope of 0.1, and, from the data, deduce the default probability threshold corresponding to this FP value. Any case with a predicted probability exceeding this threshold should be classified as a defaulter.

In Figure 6(b), we plot the slope of the graph shown in part (a). The slope function presented in part (b) is computed as the slope of the fitted curve from a nonparametric regression of TP on FP, obtained using a narrow bandwidth. A vertical line appears at the value of FP at which the slope is equal to 0.10; we see that this FP value is 1118 and we read from the data set that the corresponding default probability is 0.0813. The full confusion table at this threshold is shown in Table 3.[6]

[5] Similar methods have also been applied in the weather forecasting literature (see, for example, Jolliffe and Stephenson 2003).

[6] A confusion table is a 2×2 tabulation showing, for a given threshold, the number of true negatives (TN), false positives (FP), false negatives (FN) and true positives (TP).

If we assume that the cost of a false positive is 1, so that the cost of a false negative is 10, the total cost may be computed as follows:

$$\text{cost} = \text{FP} + 10\text{FN} = 1118 + (10 \times 312) = 4238.$$

The total cost figure thus computed is yet another measure for comparing models (see Lohmann and Ohlinger 2018a,b). Consider the in-sample results. The row of Table 2 labeled "Cost(10)" contains the total cost figures for all six models when the cost ratio is assumed to be 10. We see that, in-sample, model 3 (flex-amount) is the most preferred model on this basis, giving rise to the lowest total cost. When the cost ratio is assumed to be higher, at 15 or 20, model 6 (flex-all) gives rise to the lowest total cost. Out-of-sample, the pattern is similar. It seems that the most flexible model is the cost-minimizing model when the cost ratio is higher.

4 DISCUSSION

In this paper we have provided an application of the GAM approach to the modeling of default likelihood in a sample of personal loans. In doing so we have highlighted the major advantage of the GAM approach: the straightforwardness of estimation and testing, leading to an unambiguous strategy for model selection.

We applied the GAM estimation framework to loan default data for a sample of Korean borrowers. The continuous independent variables have nonlinear effects that became very clear when plots of the predicted default probability were obtained from the estimation results. We highlighted these plots as another attractive feature of the GAM approach. The plots conveyed interesting findings regarding the types of borrower most likely or least likely to default, and we attempted to link some of these findings to the economics literature.

We considered a range of model evaluators, covering measures of discrimination, measures of calibration and measures based on misclassification costs. These measures were obtained both in-sample and out-of-sample, and the findings are broadly similar between the two. The overall conclusion from these various evaluation routines is that the most flexible of the GAM models outperforms all other models on most criteria. It is also interesting to look at the worst-performing model for each criterion. In Table 2 we see that the worst-performing model is often model 2, which is the model that includes quadratic terms for all continuous variables but avoids fully flexible effects. Note in particular that, in out-of-sample prediction, model 2 tends to perform worse than model 1, which contains no nonlinear terms. The striking message here is that, when effects are nonlinear, the practice of simply adding quadratic terms can be counterproductive, while flexible modeling appears to be a dependable means of improving performance for most criteria.

A model evaluator that we consider to be particularly useful is one that incorporates misclassification costs. Such an evaluator must allow for the fact that the cost of an FN is higher than that of an FP. The exact cost ratio is subject to uncertainty, and for this reason we performed the cost-minimization exercise at a range of different cost ratios. The key conclusion is that, for the criterion of cost minimization, the best model is always one of the GAMs, and usually the most flexible of these: model 6.

Given the uncertainty over the cost ratio, this is clearly an area where further research is called for. In particular, while it seems reasonable to assume that all false positives incur an equal cost, it is not reasonable to expect the same of false negatives: clearly some actual defaults must be more serious, and therefore more costly, than others. A possible approach would be to assume that the cost of default is positively related to the probability of default. However, Moffatt's (2005) hurdle model provided clear evidence that the process determining the extent of a default is different than that determining whether a default occurs. This suggests that information on actual default costs would be necessary to pursue this line of enquiry.

DECLARATION OF INTEREST

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

ACKNOWLEDGEMENTS

The authors thank the Eastern Academic Research Consortium (Quantitative Social Science) for support.

REFERENCES

Abdou, H. A., Mitra, S., Fry, J., and Elamer, A. A. (2019). Would two-stage scoring models alleviate bank exposure to bad debt? *Expert Systems with Applications* **128**, 1–13 (https://doi.org/10.1016/j.eswa.2019.03.028).

Adams, N. M., and Hand, D. J. (1999). Comparing classifiers when the misallocation costs are uncertain. *Pattern Recognition* **32**(7), 1139–1147 (https://doi.org/10.1016/S0031-3203(98)00154-X).

Afifi, A. A., and Elashoff, R. M. (1966). Missing observations in multivariate statistics. I. Review of the literature. *Journal of the American Statistical Association* **61**(315), 595–604 (https://doi.org/10.1080/01621459.1966.10480891).

Bradley, M. G., Cutts, A. C., and Liu, W. (2015). Strategic mortgage default: the effect of neighborhood factors. *Real Estate Economics* **43**(2), 271–299 (https://doi.org/10.1111/1540-6229.12081).

Butaru, F., Chen, Q., Clark, B., Das, S., Lo, A. W., and Siddique, A. (2016). Risk and risk management in the credit card industry. *Journal of Banking and Finance* **72**, 218–239 (https://doi.org/10.1016/j.jbankfin.2016.07.015).

Calabrese, R., and Osmetti, S. A. (2015). Improving forecast of binary rare events data: a GAM-based approach. *Journal of Forecasting* **34**(3), 230–239 (https://doi.org/10.1002/for.2335).

Castermans, G., Martens, D., Van Gestel, T., Hamers, B., and Baesens, B. (2010). An overview and framework for PD backtesting and benchmarking. *Journal of the Operational Research Society* **61**(3), 359–373 (https://doi.org/10.1057/jors.2009.69).

Cook, J., and Ramadas, V. (2020). When to consult precision–recall curves. *Stata Journal* **20**(1), 131–148 (https://doi.org/10.1177/1536867X20909693).

De Boor, C. (2001). *A Practical Guide to Splines*, revised edn. Springer.

Efron, B. (1978). Regression and ANOVA with zero–one data: measures of residual variation. *Journal of the American Statistical Association* **73**(361), 113–121 (https://doi.org/10.1080/01621459.1978.10480013).

Fox, J. (2002). Nonparametric regression. In *An R and S-Plus Companion to Applied Regression*, Appendix. Sage Publications, London.

Greene, W. (1998). Sample selection in credit-scoring models. *Japan and the World Economy* **10**(3), 299–316 (https://doi.org/10.1016/S0922-1425(98)00030-9).

Gu, J., Li, D., and Liu, D. (2007). Bootstrap non-parametric significance test. *Nonparametric Statistics* **19**(6–8), 215–230 (https://doi.org/10.1080/10485250701734497).

Hand, D. J., and Anagnostopoulos, C. (2013). When is the area under the receiver operating characteristic curve an appropriate measure of classifier performance? *Pattern Recognition Letters* **34**(5), 492–495 (https://doi.org/10.1016/j.patrec.2012.12.004).

Hastie, T. J., and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman & Hall/CRC, Boca Raton, FL.

Hosmer, D. W., Jr., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied Logistic Regression*. Wiley (https://doi.org/10.1002/9781118548387).

Jiang, W., Nelson, A. A., and Vytlacil, E. (2014). Liar's loan? Effects of origination channel and information falsification on mortgage delinquency. *Review of Economics and Statistics* **96**(1), 1–18 (https://doi.org/10.1162/REST_a_00387).

Jolliffe, I. T., and Stephenson, D. B. (2003). *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Wiley.

Khandani, A., Kim, A., and Lo, A. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking and Finance* **34**(11), 2767–2787 (https://doi.org/10.1016/j.jbankfin.2010.06.001).

Kim, M. J., and Lee, H. S. (2010). Household financial structures by family life cycle. *Korean Journal of Community Living Science* **21**(1), 53–69.

Larsen, K. (2015). GAM: the predictive modeling silver bullet. Blog Post, July 30, *Multithreaded*, Stitch Fix. URL: https://multithreaded.stitchfix.com/blog/2015/07/30/gam.

Larsen, K. (2016). Information: data exploration with information theory (weight-of-evidence and information value). R package (Version 0.0.9). URL: https://cran.r-project.org/web/packages/Information/.

Lave, C. A. (1970). The demand for urban mass transportation. *Review of Economics and Statistics* **52**(3), 320–323 (https://doi.org/10.2307/1926301).

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

Lohmann, C., and Ohlinger, T. (2018a). Nonlinear relationships in a logistic model of default for a high-default installment portfolio. *The Journal of Credit Risk* **14**(1), 1–24 (https://doi.org/10.21314/JCR.2017.232).

Lohmann, C., and Ohlinger, T. (2018b). The total cost of misclassification in credit scoring: a comparison of generalized linear models and generalized additive models. *Journal of Forecasting* **38**(5), 375–389 (https://doi.org/10.1002/for.2545).

Medema, L., Koning, R. H., and Lensink, R. (2009). A practical approach to validating a PD model. *Journal of Banking and Finance* **33**(4), 701–708 (https://doi.org/10.1016/j.jbankfin.2008.11.007).

Moffatt, P. G. (2005). Hurdle models of loan default. *Journal of the Operational Research Society* **56**(9), 1063–1071 (https://doi.org/10.1057/palgrave.jors.2601922).

Mumpower, J. L., and McClelland, G. H. (2014). A signal detection theory analysis of racial and ethnic disproportionality in the referral and substantiation processes of the US child welfare services system. *Judgment and Decision Making* **9**(2), 114–128.

Newson, R. (2000). B-splines and splines parameterized by their values at reference points on the X-axis. *Stata Technical Bulletin* **10**(57), 20–27. URL: www.stata-press.com/journals/stbcontents/stb57.pdf.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., and Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**(1), 1–8 (https://doi.org/10.1186/1471-2105-12-77).

Saito, T., and Rehmsmeier, M. (2015). The precision–recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* **10**(3), Paper e0118432 (https://doi.org/10.1371/journal.pone.0118432).

Saito, T., and Rehmsmeier, M. (2017). Precrec: fast and accurate precision–recall and ROC curve calculations in R. *Bioinformatics* **33**(1), 145–147 (https://doi.org/10.1093/bioinformatics/btw570).

Schechtman, E., and Schechtman, G. (2019). The relationship between Gini terminology and the ROC curve. *Metron* **77**(3), 171–178 (https://doi.org/10.1007/s40300-019-00160-7).

Stone, C. J. (1985). Additive regression and other nonparametric models. *Annals of Statistics* **13**(2), 689–705 (https://doi.org/10.1214/aos/1176349548).

Taylan, P., Weber, G. W., and Beck, A. (2007). New approaches to regression by generalized additive models and continuous optimization for modern applications in finance, science and technology. *Optimization* **56**(5–6), 675–698 (https://doi.org/10.1080/02331930701618740).

Thomas, L., Crook, J., and Edelman, D. (2017). *Credit Scoring and Its Applications*. SIAM, Philadelphia, PA (https://doi.org/10.1137/1.9781611974560).

Yu, L., Wang, S., and Lai, K. K. (2009). An intelligent-agent-based fuzzy group decision making model for financial multicriteria decision support: the case of credit scoring. *European Journal of Operational Research* **195**(3), 942–959 (https://doi.org/10.1016/j.ejor.2007.11.025).