GENERALISED ADDITIVE MODELLING OF THE CREDIT RISK OF KOREAN PERSONAL BANK LOANS

Young-Ah KIM
Essex Business School
10 Elmer Approach,
Southend-On-Sea SS1 1LW
United Kingdom

Peter G MOFFATT[1]
School of Economics
University of East Anglia
Norwich NR4 7TJ
United Kingdom

Simon A PETERS
School of Social Sciences
University of Manchester
Oxford Road, Manchester M13 9PL
United Kingdom

5 OCTOBER 2021

Running Header: Credit Risk of Korean Personal Loans

Number of words (excluding references): 6450

Figure and Table Legends:

Key words and phrases: Generalised additive models; B-spline; credit scoring; loan defaults; signal detection theory; mis-classification costs.

<u>Declaration of Interest statement</u>

The authors thank the EARC QSS (Eastern Academic Research Consortium Quantitative Social Science) for support. The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

[1] Corresponding Author: Peter G. Moffatt, School of Economics, University of East Anglia, Norwich, NR4 7TJ, United Kingdom. Email: p.moffatt@uea.ac.uk

GENERALISED ADDITIVE MODELLING OF THE CREDIT RISK OF KOREAN PERSONAL BANK LOANS

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 non-linear impact of loan and borrower characteristics. In particular, the likelihood of default is high for both low income borrowers as well as high income borrowers. Our results are robust to a range of different tests, and highlight the usefulness of the GAM framework, especially the graphical presentation of non-linearities.

GENERALISED ADDITIVE MODELLING OF THE CREDIT RISK OF KOREAN PERSONAL BANK LOANS

1. Introduction

Models for binary outcomes, such as logit and probit, are popular tools in the modelling of loan defaults (see for example Greene, 1998; Thomas et al., 2017). Such models are usually classified under the heading of Generalised Linear Models (GLMs). In this paper, we extend the GLM framework by using Generalized Additive Modelling (GAM, Hastie and Tibshirani, 1990). The essence of the approach is that each individual effect is modelled separately using a non-parametric smoothing procedure, allowing non-linear 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 non-linear effects to be specified before estimation, the GAM approach finds the pattern through estimation, and is hence not constrained by prior beliefs. Clearly the greater flexibility of GAM's has the potential to bring about improvements in terms of superior ability to discriminate between defaulters and non-defaulters, greater accuracy in predicted default probabilities, and lower mis-classification 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. Recent applications of these methods to credit default data include Khandani et al. (2010), Butaru et al. (2016), and Abdou at al. (2019). A recent survey of such methods is provided by Lessman et al. (2015). A potential problem with all of these approaches (including GAMs) is that of over-fitting, and for this reason, there is clear need to play 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 is both straightforward and transparent in the context of GAMs, leading to a reliable and unambiguous strategy for model selection. A more important reason relates to interpretability: as will be demonstrated in this paper, the structure of the 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 the effect in ways that appeal to non-specialists. A further reason relates to 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 tradeoff.

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

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 non-linear effects are characteristics of individual borrowers such as age, income, and amount borrowed, and we will focus on these non-linear effects in the interpretation of the model results. We will highlight the graphical presentations of these non-linear 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 B-spline smoother (de Boor, 2001). The B-spline procedure offers an attractive compromise between polynomial regression and kernel regression. The former provides 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 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 non-parametric technique that can be performed as a (generalised) 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 non-standard distributions and/or resampling methods (see, e.g. 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 judgements on whether a predictor should be represented flexibly at all, in preference to assuming a linear effect.

We estimate a number of models, both GLMs and GAMs, varying in flexibility. We also use a variety of approaches to evaluate the predictive performance of the estimated models. A very useful recent survey of these techniques is provided by Lessmann et al. (2015). We start by applying the well-established receiver operating characteristics (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 which 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 mis-classification costs, and we use total mis-classification 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 from applying the GAM framework to the data set, and also apply a range of model evaluation techniques in order to compare the the performance of GAMs to that of less flexible models. Section 4 provides a summary of the findings.

2. **Modelling Strategy and Data**

2.1 **Generalised Additive Models (GAMs)**

Traditional regression models frequently fail for the simple reason that the effects of interest are often non-linear. To characterise 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 non-continuous (e.g. binary) outcomes, further generality is required. Hastie and Tibshirani (1990) introduced the framework of Generalized Additive Models (GAMs). These 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, and if we assume a logit link function, the model takes the following form:

$$P(y_i = 1|x_{1i}, \cdots, x_{mi}) = \Lambda\left(\beta_0 + \sum_{j=1}^{m} f_j\left(x_{ji}\right)\right) \qquad (1)$$

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

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", $\xi_1, \ldots, \xi_K$, and 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 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 2nd and lower order discontinuities, but not higher".

The more knots are 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 can be used. Another choice that needs to be made is the positioning of the knots. The approach adoped 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", as used in the model of later sections. There are five knots, and therefore seven basis functions used in the regression. The basis functions may computed using a method due to Newson (2000).

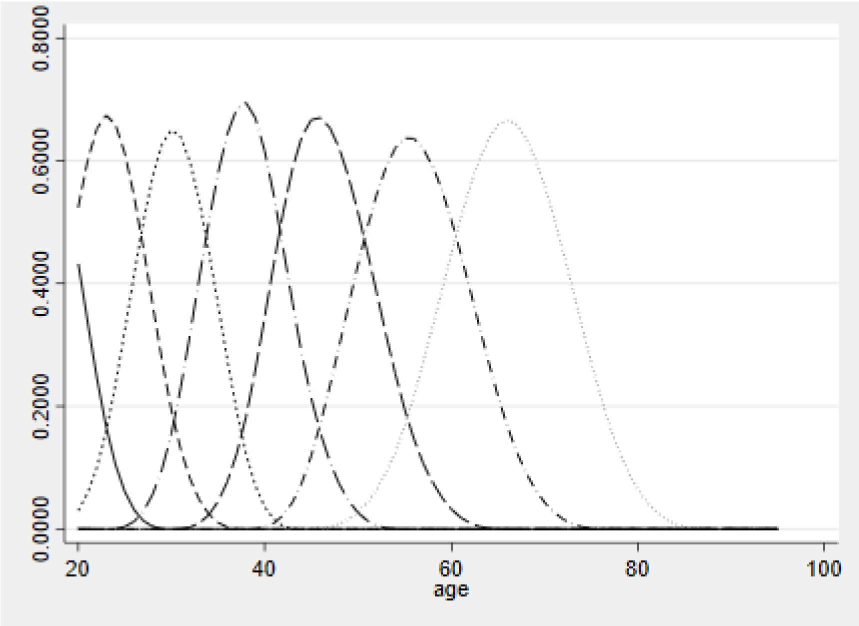

Figure 1: basis functions for variable age; knots at 23,30,38,44,55.

If the basis functions to be used in the B-spline regression are $B_1(.) \cdots B_{K+2}(.)$, then the piecewise cubic functions $f_j(.)$ appearing in (1) may be expressed as:

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

It is important that (2) does not contain an intercept. This is necessary for the model intercept (β_0 in (1)) to be identified.

Note also that (2) would lead to a fully general GAM in the sense that all *m* of the predictors are being assumed to have flexible effects. As noted in the discussion following (1), there are strong reasons for *not* modelling the effect of every predictor in accordance with (2).

As mentioned in Section 1, an understated advantage of the "GAM with B-spline" approach is that it can be estimated with a linear regression. Although the regression coefficients on the basis functions (i.e. estimates of the γ_{jk}'s in (2)) are themselves hard to interpret, it is a straightforward matter to perform regression-based tests (e.g. for predictor *j*, a joint test of H_0: γ_{j1}=0;… $\gamma_{j,K+2}$=0) in order firstly to test for the presence of non-linear effects of predictors, and secondly to 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 of the 48 similar studies considered by Lessman et al. (2015), which was 6,167.

Of the 64,579 borrowers, 32,534 fail to report income. Since income one of the key determinants of default, the estimation sample consists only of 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 is performed of income 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 the bar is being raised for all models, so that 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 Tables A1 and A2 of the Appendix, with variable definitions provided in Table A3. In Figure A1, we present histograms showing the distributions of the three key continuous independent variables entering 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.

Purpose-of-loan is divided into five categories: unspecified, property, financial, living expenses, and other. Definitions of these categories, in terms of lists of sub-categories, are provided in Table A3. Table A4 shows descriptive statistics of loan amount separately by 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 property loans account for a small proportion of the total, and it is the case that 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 31 December 2012: "default" is defined as any sort of failure to meet the obligations of the loan during the time period 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 routinely included in estimation.

Operationally, banks in Korea aim to achieve an overall default rate for individual loans of between 1.5% and 2.0%.[2] As seen from the first row of Table A1, the overall default rate in our estimation sample is marginally below 1.5%. However, note from Table A2 that the default rate in the test sample is a much higher 4.3%. Remembering that the test sample consists of borrowers from whom income is not observed, this difference clearly suggests that missing income is a strong risk factor, and also presents a challenge for out-of-sample prediction.

[2] See https://fsckorea.wordpress.com/2012/06/11/monitoring-the-loan-default-rate-of-korean-banks/

3. Results

3.1 Model Estimates

The dependent variable in our analysis is 1 if the borrower defaulted and zero otherwise. The predictors are: log(amount);[3] log(income); age in years; male (1 if male, 0 if female); married (1 if married, 0 otherwise); number of dependants; four dummy variables for purpose of loan:[4] property, financial, living, other (base case: unspecified purpose). There is also a variable representing the duration of the loan, measured in years (duration). 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 non-linear 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 non-linearities whether (e.g.) income or log(income) is being considered. WoE scores for categorised 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 Information package of Larsen (2016). None of the 3 WoE score vectors show a monotonic change as the category values increase. This non-monotonic behaviour is usually interpreted as an indication that these variables may affect the default probability in a non-linear fashion. A further output of WoE analysis is the Information Value (IV), which acts as a summary of the WoE over all categories. The IV's of all variables are shown in the final column of Table A1 of the Appendix. According to a rule-of-thumb,[5] an IV less than 0.02 indicates a useless predictor, one less than 0.10 indicates a weak predictor, one between 0.10 and 0.3 indicates a "medium" predictor, and one 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.

[3] We acknowledge the possibility that there may be an element of endogeneity in the variable "amount" since the size of the approved loan is likely to be determined partly by the perceived credit-worthiness of the borrower. The broadly negative effect of this variable seen in both of Figures 2 and 3 is consistent with this possibility.
[4] See Table A3 for further disaggregation of these purpose categories.
[5] https://www.listendata.com/2015/03/weight-of-evidence-woe-and-information.html

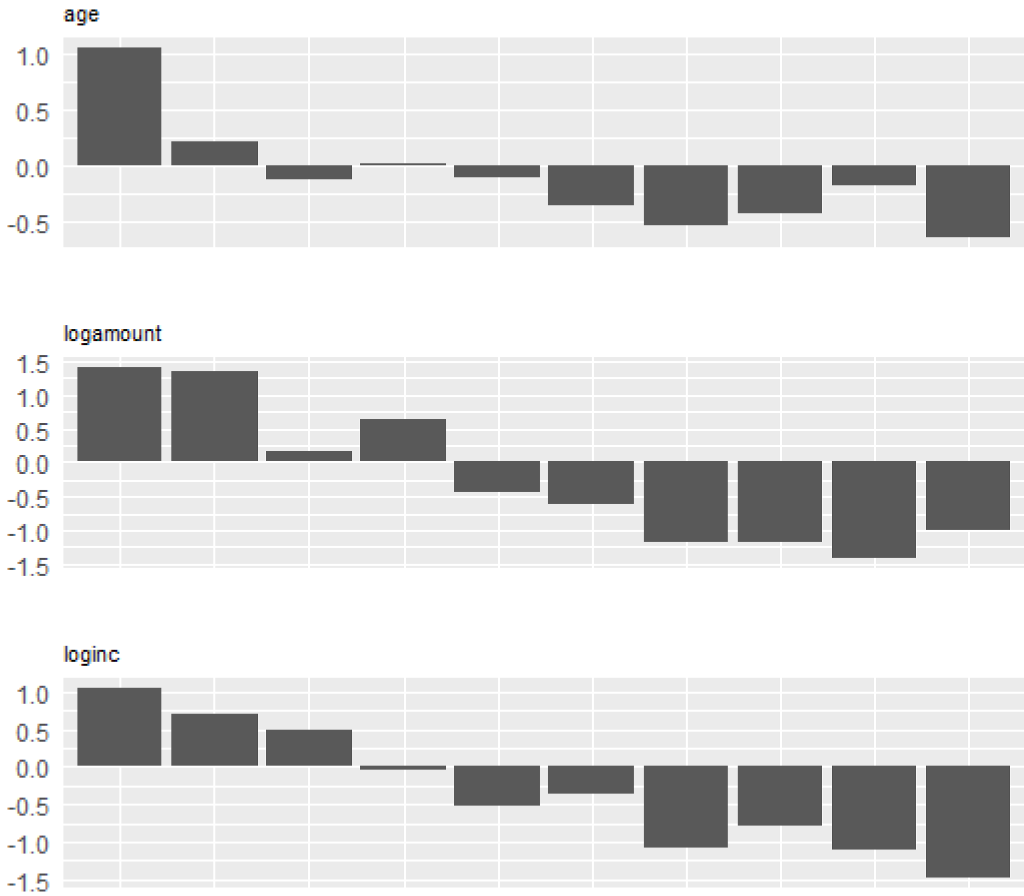

Figure 2: WoE Score Vectors for Age, Log (Amount) and Log (Income)

Six binary logit models have been estimated, and the results are reported in Table 1. Model 1 (Baseline) is a simple logit model with linear effects only. Model 2 (Quadratic) is a simple logit model with quadratic effects assumed for the three continuous predictors (log(amount), log(income), and age). On the basis of Akaike's 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 default probability, there is no evidence of a quadratic effect of log(amount).

	(1) Baseline	(2) Quadratic	(3) Flex-Amount	(4) Flex-Income	(5) Flex-Age	(6) Flex-All
Logamount	-0.363***	-0.802		-0.811	-0.847	
	(-5.96)	(-1.54)		(-1.59)	(-1.63)	
Logamount2		0.0158		0.0167	0.0171	
		(0.95)		(1.02)	(1.03)	
Logincome	-0.188***	0.909**	0.785*		0.969**	
	(-5.82)	(2.79)	(2.41)		(2.95)	
Logincome2		-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)		
Age2		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)
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)
$\gamma_{logamount,1}$			0.186			0.187
			(0.32)			(0.33)
$\gamma_{logamount,2}$			0.445			0.422
			(1.04)			(0.98)
$\gamma_{logamount,3}$			-0.733			-0.707
			(-1.62)			(-1.56)
$\gamma_{logamount,4}$			-0.367			-0.340
			(-0.84)			(-0.77)
$\gamma_{logamount,5}$			-1.258*			-1.180*
			(-2.28)			(-2.13)
$\gamma_{logamount,6}$			-1.527*			-1.535*
			(-2.05)			(-2.05)
$\gamma_{logamount,7}$			-0.369			-0.338
			(-0.37)			(-0.34)
$\gamma_{logincome,1}$				0.0917		0.139
				(0.14)		(0.21)
$\gamma_{logincome,2}$				-0.0782		-0.0442
				(-0.19)		(-0.11)
$\gamma_{logincome,3}$				-0.331		-0.318
				(-0.91)		(-0.87)

$\gamma_{logincome,4}$				-1.106^{**}		-0.983^{**}
				(-3.19)		(-2.80)
$\gamma_{logincome,5}$				-0.879^{*}		-0.811
				(-1.97)		(-1.81)
$\gamma_{logincome,6}$				-1.670^{**}		-1.599^{*}
				(-2.68)		(-2.53)
$\gamma_{logincome,7}$				-1.336		-1.123
				(-1.16)		(-0.97)
$\gamma_{age,1}$					7.189^{*}	7.288^{*}
					(2.45)	(2.45)
$\gamma_{age,2}$					-0.120	-0.184
					(-0.08)	(-0.12)
$\gamma_{age,3}$					0.963	0.953
					(0.65)	(0.63)
$\gamma_{age,4}$					-0.546	-0.502
					(-0.38)	(-0.34)
$\gamma_{age,5}$					-0.00999	0.0297
					(-0.01)	(0.02)
$\gamma_{age,6}$					-0.200	-0.133
					(-0.15)	(-0.10)
$\gamma_{age,7}$					-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	32045	32045	32045	32045	32045	32045
k	12	15	20	20	20	30
AIC	4294.5	4276.6	4254.4	4274.2	4275.0	4253.0

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1: Results of six binary logit models of loan default. t-statistics in parentheses. Model 1: linear effects only. Model 2: quadratic effects. Model 3: GAM with flexible effect of log(amount) (knots: 14.9,15.4,16.1,16.8,18.1). Model 4: GAM with flexible effect of log(income) (knots: 15.9,16.8,17.3,17.8,18.2). Model 5: GAM with flexible effect of age (knots: 23,30,38,44,55). Model 6: GAM with flexible effects of all three. LogL is maximised log-likelihood. AIC = 2(k-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={logamount, logincome, age} k=1,…,7 are the coefficients on the basis functions used to obtain the B-spline (see Equation (2)).

Models 3-6 are GAMs. Model 3, 4, and 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 non-linear 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 AIC, this model is the best performer 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 a relatively straightforward matter to use the estimated coefficients to generate predicted probability curves, with confidence bands, 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 for the first two plots, 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.

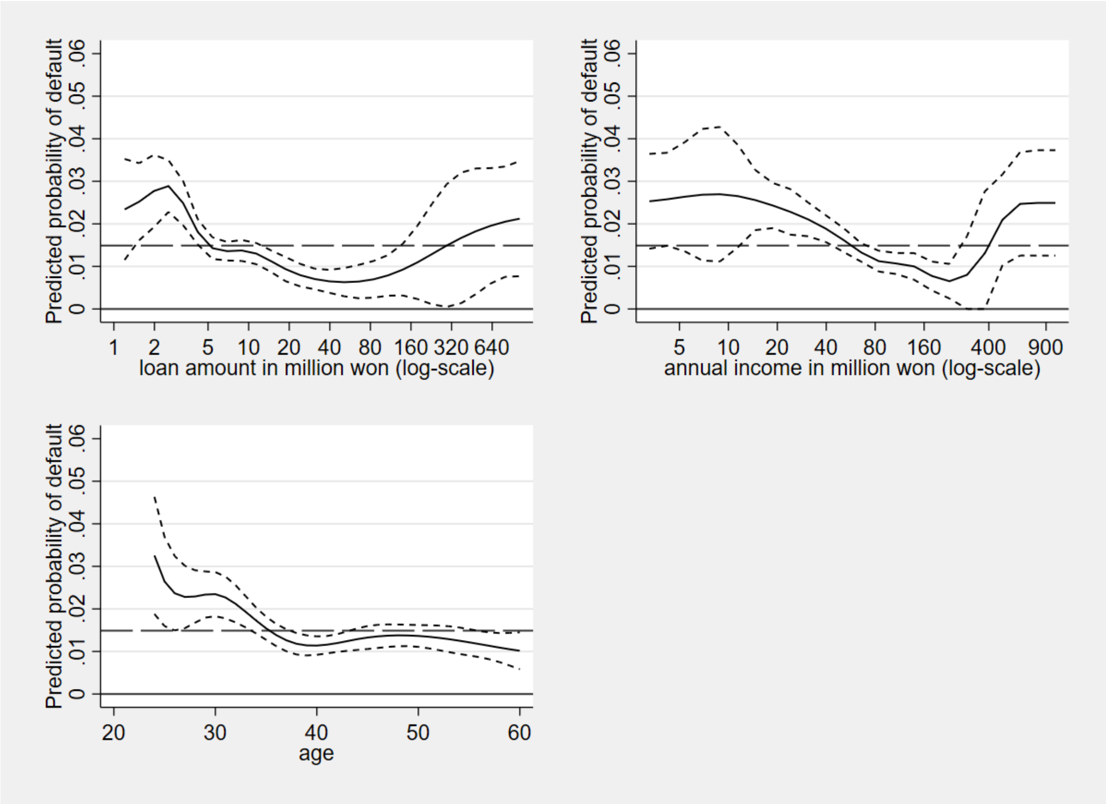

Figure 3: Predicted probability of default against amount (top-left panel; log-scale) income (top-right panel; log-scale) and age (bottom panel). Other variables set to means. Solid curves represent predicted probabilities; short-dashed curves represent 95% confidence bands; long-dashed lines drawn at sample proportion of defaults.

In the cases of log(amount) and log(income), the most striking feature of the plots is the pronounced uptick at the upper end of the scale. Possible explanations for these upticks are available from the Economics literature. For example, it is conceivable that loans offered to borrowers with the highest reported income may be “liar’s loans” (Jiang et al., 2014); that is, 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 (e.g. Bradley et al., 2015) that strategic default is strongly associated with larger loan amounts, hence 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, the third plot in Figure 3 also shows hints that the effect of age takes the form of a downward step function, levelling off at particular stages of the life cycle, namely late-twenties and early-forties. These age-ranges correspond to definitive stages of 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 number of dependants are apparently unimportant. Purpose of loan is important, with those borrowing for "financial" purposes" (which includes the sub-categories: Business; Investment; Repayment of other loans; Repayment of credit card) and "living expenses" both 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, firstly because the bank's credit scoring algorithm is unknown to us, and secondly 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 then compared using Akaike's Information Criterion (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 predictive performance in 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 minimisation. 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 the test sample consisting of 32,534 observations (with missing income imputed).

"Discrimination" refers to the ability of a model to separate defaulters from non-defaulters. 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 Rehsmeier, 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 Receiver Operating Characteristics (ROC) curve. To assess the discriminatory performance of a model, 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, one 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-degree line. The higher above the 45-degree line the ROC curve lies, the better the disciminatory performance of the model, since this implies that TPR is rising faster than FPR when the threshold is lowered. A model that predicts perfectly (with the predicted probability perfectly separating the two groups) will produce a ROC curve consisting of a vertical 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 (AUC_{ROC}). According to Hosmer et al. (2013), AUC_{ROC} of 0.5 indicates no discrimination, AUC_{ROC} between 0.7 and 0.8 indicates acceptable discrimination, AUC_{ROC} between 0.8 and 0.9 indicates excellent discrimination, and AUC_{ROC} greater than 0.9 indicates outstanding discrimination. As noted by Adams and Hand (1999), AUC_{ROC} has the intuitive interpretation of representing the probability that a randomly chosen default case will have been ranked higher by the model than a randomly chosen non-default case. There is also a close link between AUC_{ROC} and the Gini coefficient that is widely used in the Economics literature as a measure of inequality (see Schechtman and Schechtman, 2019).

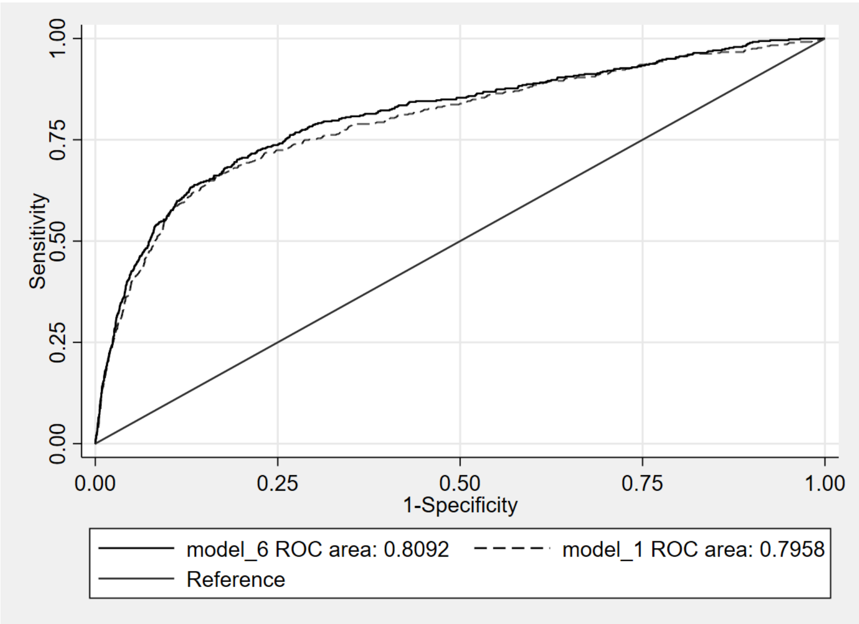

Figure 4: ROC curves from Models 1 and 6. AUC_{ROC} = 0.796, 0.809 respectively.

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 (AUC_{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 AUC's 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 AUC's obtained using the same data set. To address this issue, we follow Robin et al. (2011) by bootstrapping the *difference* between the two AUC's 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 AUC_{ROC}, and all of the other model selection criteria discussed in this section, for all six of the models estimated in Section 3.1, applied both in-sample and out-of-sample. The row below the AUC_{ROC} contains the p-values for (one-sided) tests of 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.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
	Baseline	Quadratic	Flex-Amount	Flex-Income	Flex-Age	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*
IN-SAMPLE:						
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*
OUT-OF-SAMPLE:						
AUC_{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	-
AUC_{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)	11374.4	11372.5	11362.1	11311.0*	11474.1†	11363.2
Cost(15)	14275.4†	14281.8	14182.2	14251.5	14220.8	14167.8*
Cost(20)	16,566	16,612†	16,556	16,599	16,365	16,230*

Table 2: Model selection criteria for the six models estimated in Section 3.1. Top panel reproduces key model selection criteria reported in Table A1. Second panel reports measures of in-sample predictive performance (estimation sample of 32,045 used for both estimation and prediction). Third panel reports measures of out-of-sample predictive performance (estimation sample of 32,045 observations used for estimation; test sample of 32,534 used for prediction). AUC_{ROC} is area under ROC curve; AUC_{PRC} is area under Precision Recall Curve. P-value appearing in row below AUC_{ROC} (resp. AUC_{PRC}) is for the (one-tailed) test of the difference of the model's AUC_{ROC} (resp. AUC_{PRC}) from that of Model 6 (201 Bootstrap replications). MSE is mean-sqaued error (a.k.a. Brier Score). R^2_{LE} is Lave-Effron R-squared. Cost(.) is the total mis-classification cost when the number in parentheses is the cost ratio, and the cost of a FP is 1. Stars (*) indicate best-performing model on each criterion; daggers (†) indicate worst-performing model.

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 might be perceived as a suggestion of over-fitting of the other models, but this is not seen anywhere else in the results.

It has been argued by Saito and Rehmsmeier (2015, 2017) that the ROC approach used above may be misleading in the presence of an unbalanced data set. According to Saito and Rehmsmeier (2015, p.1), "visual interpretability of ROC plots in the context of imbalanced datasets 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. These authors have suggested that the Precision Recall Curve (PRC) is used in addition to ROC in the presence of unbalanced data. The PRC is a plot of positive predicted value (PPV) against TPR, where PPV is defined as the number of true positives as a proportion of the total number of positives. PPV is the proportion of the observations predicted by the model to be defaulters who are true defaulters, while TPR is the proportion of defaulters who are true defaulters. The PRC therefore represents the trade-off between these two proportions. Again, the area under this curve, AUC_{PRC}, will provide a numerical measure of discriminatory performance. Both PRC and 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. AUC_{PRC}'s 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 earlier for the comparison of AUC_{ROC}'s. 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.

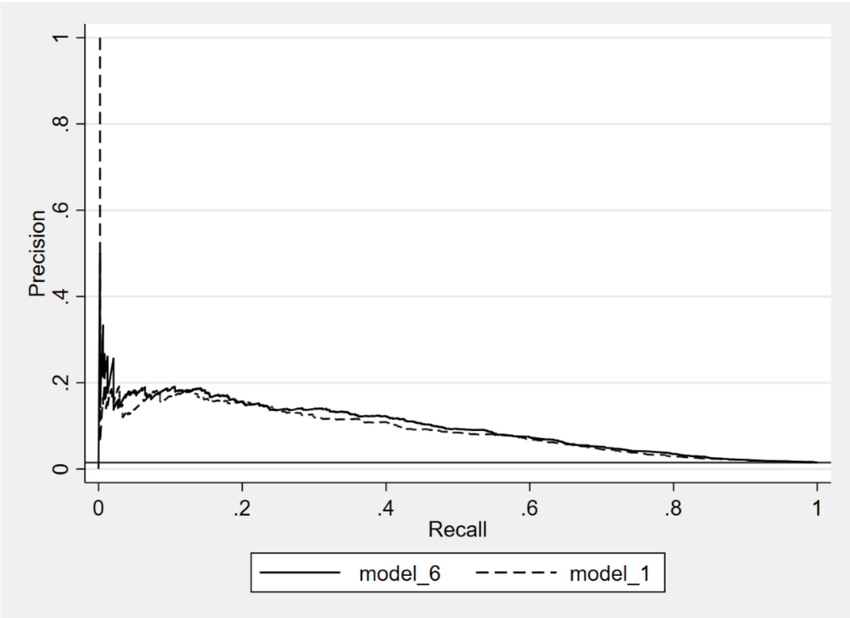

Figure 5: PRC's from Models 1 and 6. AUC_{PRC} = 0.088, 0.097 for Models 1 and 6 respectively.

The second type of measure of predictive performance come under the heading of "calibration". "Calibration" refers to the closeness with which predicted probabilities correspond to actual outcomes. Calibration is clearly important if accurate estimation of default probabilities is what is required.

One obvious calibration measure is 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, R^2_{LE} (Lave, 1970; Efron, 1978). For each of the models estimated in Section 3.1, we report both MSE and 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 over-fitting is not an issue.

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 how to select a classification rule, and then how to assess the performance of the model conditional on the chosen rule.

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

Adams and Hand (1999) emphasise the uncertainty surrounding the mis-classification cost ratio, but, on the basis of discussions with banking domain experts, 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. Lessman 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 in which all loans are approved.

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

---

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

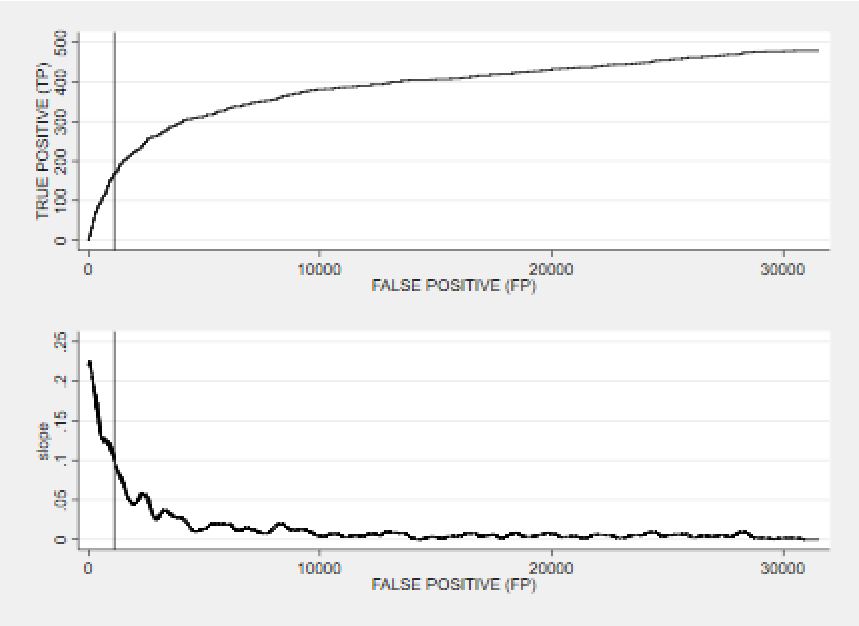

Figure 6: Upper panel: Plot of number of true positives (TP) against number of false positives (FP), from Model (6); Lower panel: slope of graph shown in top panel. Vertical line appears at value of FP at which slope = 0.10.

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

In the lower panel of Figure 6, we plot the slope of the graph shown in the upper panel. The slope function presented in the lower panel of Figure 6 is computed as the slope of the fitted curve from a non-parametric regression of TP on FP obtained using a narrow bandwidth. A vertical line appears at value of FP at which slope = 0.10, and we see that this value of FP is 1118, and we read from the data set that the corresponding default probability is 0.0813. The full confusion table[7] at this threshold is shown in Table 3.

	P < 0.0758	***P ≥ 0.0758***
Non-default	TN = 30,448	FP = 1,118
Default	FN = 312	TP = 167

Table 3: Confusion table for threshold 0.0758 (Optimal for Model 6 assuming cost ratio = 10). TN = number of true negatives; FN = false negatives; TP = true positives; FP = false positives. Total sample size: 32,045.

[7] 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 FP is 1, so that the cost of a FN is 10, the total cost may be computed as follows:

$$cost = FP + 10FN = 1118 + 10 \times 312 = \underline{4238}$$

The total cost figure thus computed is yet another measure for comparing models (see Lohmann and Ohlinger, 2018a, 2018b). Consider the in-sample results. The row of Table 2 named 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 and 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-minimising 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 have applied the estimation framework to loan default data on a sample of Korean borrowers. The continuous independent variables have non-linear effects which became very clear when plots of predicted default probability were obtained from the estimation results. We have 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 have considered a range of model evaluators, covering measures of discrimination, measures of calibration, and measures based on mis-classification costs. These measures have been 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 out-performs all other models on most criteria. It is also interesting to look at the worst-performing model on each criterion. In Table 2, we see that the worst-performing model is often Model 2, which is the model that includes quadratic terms on 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 non-linear terms. The striking message here is that when effects are non-linear, the practice of simply adding quadratic terms can be counter-productive, while flexible modeling appears to be a dependable means of improving performance on 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 a FN is higher than that of a FP. The exact cost ratio is subject to uncertainty, and for this reason we have performed the cost minimisation exercise at a range of different cost ratios. The key conclusion is that on the criterion of cost minimisation the best model is always one of the GAMs, and usually the most flexible of these, Model 6 (Flex-All).

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 FPs incur an equal cost, it is not reasonable to expect the same of FNs – 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 from that determining whether a default occurs. This suggests that information on actual default costs would be necessary to pursue this line of enquiry.

References

Abdou, H. A., Mitra, S., Fry, J., & Elamer, A. A. (2019). Would two-stage scoring models alleviate bank exposure to bad debt?. *Expert Systems with Applications*, *128*, 1-13.

Adams, N. M., & Hand, D. J. (1999). Comparing classifiers when the misallocation costs are uncertain. *Pattern Recognition*, *32*(7), 1139-1147.

Afifi, A. A., & Elashoff, R. M. (1966). Missing observations in multivariate statistics I. Review of the literature. *Journal of the American Statistical Association*, *61*(315), 595-604.

Bradley, M. G., Cutts, A. C., & Liu, W. (2015). Strategic mortgage default: The effect of neighborhood factors. *Real Estate Economics*, *43*(2), 271-299.

Butaru, F., Q. Chen, B. Clark, S. Das, A. Lo, A. Siddique, 2016, Risk and risk management in the credit card industry, *Journal of Banking and Finance*, 72, 218-239.

Calabrese, R., & Osmetti, S. A. (2015). Improving Forecast of Binary Rare Events Data: A GAM-Based Approach, *Journal of Forecasting*, 34, 230-239.

Castermans, G., Martens, D., Van Gestel, T., Hamers, B., & Baesens, B. (2010). An overview and framework for PD backtesting and benchmarking. *Journal of the Operational Research Society*, *61*(3), 359-373.

Cook, J., & Ramadas, V. (2020). When to consult precision-recall curves. *The Stata Journal*, *20*(1), 131-148.

De Boor, C. (2001), *A Practical Guide to Splines (Revised Edition)*, New York: 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.

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

Greene, W. (1998). Sample selection in credit-scoring models. *Japan and the World Economy*, *10*(3), 299-316.

Gu, J. Li, D. & Liu, D. (2007). Bootstrap non-parametric significance test. *Nonparametric Statistics*, 19(6-8), 215-230.

Hand, D. J., & 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.

Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized Additive Models*, New York: Chapman and Hall.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Jiang, W., Nelson, A. A., & 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.

Jolliffe, I. T., & Stephenson, D. B. (Eds.). (2003). *Forecast verification: a practitioner's guide in atmospheric science*. John Wiley & Sons.

Khandani, A., A. Kim, A. Lo, 2010, Consumer credit-risk models via machine-learning algorithms, *Journal of Banking and Finance* 34, 2767-2787

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] Retrieved from GAM: The Predictive Modeling Silver Bullet | Stitch Fix Technology – Multithreaded [Accessed August 5 2021].

Larsen, K. (2016), Information: Data Exploration with Information Theory (Weight-of Evidence and Information Value). R package version 0.0.9. https://CRAN.R-project.org/package=Information

Lave, C. A. (1970). The demand for urban mass transportation. *The Review of Economics and Statistics*, 320-323.

Lessmann, S., Baesens, B., Seow, H. V., & 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.

Lohmann, C., & Ohlinger, T. (2018a). Nonlinear relationships in a logistic model of default for a high-default installment portfolio. *Journal of Credit Risk*, 14(1), 1-24.

Lohmann, C., & Ohlinger, T. (2018b). The total cost of misclassification in credit scoring: A comparison of generalized linear models and generalized additive models. *Journal of Forecasting*. DOI:10.1002/for.2545.

Medema, L., Koning, R. H., & Lensink, R. (2009). A practical approach to validating a PD model. *Journal of Banking & Finance*, *33*(4), 701-708.

Moffatt, P. G. (2005). Hurdle models of loan default. *Journal of the Operational Research Society*, *56*(9), 1063-1071.

Mumpower, J. L., & 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.

Newson, R. (2000). sg151: B-splines and splines parameterized by their values at reference points on the X-axis. *Stata Technical Bulletin* 57: 20-27.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC bioinformatics*, *12*(1), 1-8.

Saito, T., & 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), e0118432.

Saito, T., & Rehmsmeier, M. (2017). Precrec: fast and accurate precision–recall and ROC curve calculations in R. *Bioinformatics*, *33*(1), 145-147.

Schechtman, E., & Schechtman, G. (2019). The relationship between Gini terminology and the ROC curve. *Metron*, *77*(3), 171-178.

StataCorp., (2019). *Stata Statistical Software: Release 16*. College Station, TX: StataCorp LLC.

Stone, C.J. (1985). Additive regression and other nonparametric models. *The Annals of Statistics*, 13(2), 689–705.

Taylan P., Weber G.W. & Beck, A. (2007). New Approaches to Regression by Generalized Additive Models and Continuous Optimization for Modern Applications in Finance, Science and Technology. *Optimization* 56, 675-698.

Thomas, L., Crook, J., & Edelman, D. (2017). *Credit scoring and its applications* Philadelphia : Society for Industrial and Applied Mathematics

Yu, L., Wang, S., & 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.