A Mixture Model for Credit Card Exposure at Default using the GAMLSS Framework ⋆

Suttisak Wattanawongwan[a,∗], Christophe Mues[b], Ramin Okhrati[c], Taufiq Choudhry[b], Mee Chi So[b]

[a]*School of Mathematical Sciences, University of Southampton, Highfield, Southampton, SO17 1BJ, UK*
[b]*Southampton Business School, University of Southampton, Highfield, Southampton, SO17 1BJ, UK*
[c]*Institute of Finance and Technology, University College London, London, WC1E 6BT, UK*

Abstract

The Basel II and III Accords propose estimating the Credit Conversion Factor (CCF) to model Exposure At Default (EAD) for credit cards and other forms of revolving credit. Alternatively, recent work has suggested it may be beneficial to predict the EAD directly, i.e. modelling the balance as a function of a series of risk drivers. In this paper, we propose a novel approach combining two ideas proposed in the literature and test its effectiveness using a large dataset of credit card defaults not previously used in the EAD literature. We predict EAD by fitting a regression model using the Generalized Additive Models for Location, Scale and Shape (GAMLSS) framework. We conjecture that EAD level and the risk drivers of its mean and dispersion parameters could substantially differ between the debtors who hit the credit limit (i.e. "maxed out" their cards) prior to default and those who did not, and thus implement a mixture model conditioning on these two respective scenarios. In addition to identifying the most significant explanatory variables for each model component, our analysis suggests that predictive accuracy is improved, both by using GAMLSS (and its ability to incorporate non-linear effects) as well as by introducing the mixture component.

Keywords:
Risk analysis, Basel accords, Credit cards, Exposure At Default, Generalized Additive Models

1. Introduction

The Basel regulatory accords have set out risk-sensitive regulatory capital requirements stipulating the minimum level of capital that banks must hold as a function of various types of risk. Under the Advanced Internal Rating Based (A-IRB) approach, authorised banks are permitted to use their own methods to calculate three parameters that are central to one such source of risk — credit risk. These are: Probability of Default (PD), Loss Given Default (LGD) and Exposure At Default (EAD). In retail credit risk, PD and LGD have thus far received the bulk of attention by credit risk researchers, whilst EAD has been studied far less extensively. This paper is motivated by this fact and aims to close such gap by focusing on EAD modelling.

EAD is defined as the outstanding debt at the time of default and measures the potential loss the bank would face in the absence of any further repayments. The A-IRB approach requires producing suitable EAD estimates for all loans that are not yet in default. For some types of loans, those estimates can be relatively straightforward; for example, the EAD for term loans, such as residential mortgages and personal loans, could be inferred simply from the current exposure amount plus potential subsequent interest and fees (Witzany, 2011). In contrast, for revolving retail exposures, such as credit cards and overdrafts, the estimation is more complex as customers are allowed to draw up to a specified limit and can repay any amount at any time (as long as the minimum level is met). As a result, each borrower's account balance may change substantially in the run-up to default and using the current balance may severely underestimate the true exposure risk. For these types of credit, the Basel Accords have suggested estimating a Credit Conversion Factor (CCF),

⋆Declaration of interest: none.
∗Corresponding author
Email address: S.Wattanawongwan@soton.ac.uk (Suttisak Wattanawongwan)

which is usually defined as the proportion of the undrawn amount (i.e. credit limit minus drawn amount) that will be drawn by the time of default. This CCF should reflect the likelihood of additional drawings between estimation and default time. From the predicted CCF, the estimated EAD then follows as:

$$\text{EAD} = \text{Current drawn amount} + (\text{CCF} \times \text{Current undrawn amount}).$$

Even though statistical methods to estimate the CCF have been proposed, several drawbacks were soon identified. For example, the CCF distribution is highly bimodal, estimates must be restricted to the [0,1] range, and models may struggle to cope with the contracting denominator when the current drawn amount is already close to the limit. Therefore, in the literature, alternative methods have been suggested to avoid the undesired properties of CCF models, which include predicting EAD directly (Tong et al., 2016).

In this paper, we focus on EAD modelling for credit cards, which has received limited attention in the literature. Most of the studies on EAD modelling have thus far focused on corporate credit, whilst fewer address retail customers (Gürtler et al., 2018). This is partly explained by the greater availability of public data on the corporate sector and by the fact that the financial status and health of corporate customers could be inspected from share and market-traded products (Leow and Mues, 2012), enabling easier access to data. Nonetheless, credit cards make up the largest share of revolving retail credit for most A-IRB banks and contribute the largest number of defaults compared to other revolving line products (Qi, 2009). This should contribute sufficiently large information about the characteristics of defaulted accounts to enable statistical modelling.

To avoid the problems associated with CCF estimation, we choose the EAD amount itself as the response variable. This choice, however, poses other challenges. For example, the observed value range of realised EAD levels could be very wide and thus difficult to capture statistically (Yang and Tkachenko, 2012). To cope with its right-skewness, Tong et al. (2016) therefore proposed a gamma distribution for (non-zero) EAD and built a direct EAD model under the Generalized Additive Models for Location, Scale and Shape (GAMLSS) framework (Stasinopoulos et al., 2017), which was shown to outperform several benchmark models (including for CCF) on a dataset from a UK lender. In this paper, we take a similar approach but we further extend it by distinguishing between two subgroups of credit card borrowers — those whose balance hit the limit at least once in the run-up to default, versus those who never maxed out their card over that same outcome period —, introducing two mixture components to our models. The rationale for doing so is that we hypothesise that not just the EAD but also its risk drivers (and that of its dispersion) could differ substantially between the two groups. A similar mixture element was previously proposed by Leow and Crook (2016), along with their panel models for card balance (and limit), but besides us using a different modelling framework applied to (cross-sectional) default cohort data, our approach differs from theirs in that we allow for non-parametric terms, and nor do we assume that the balance of maxed-out accounts has to match the limit value exactly.

To empirically validate the effectiveness of the GAMLSS model (versus OLS), the proposed mixture approach, and its combined application, we construct a set of benchmark models against which we compare the predictive performance of our newly proposed model. All models are fitted using a large dataset of credit card defaults from a large Asian lender, which has not been previously used in the EAD literature.

To summarise, the contributions of our new model and analysis are that we: (1) estimate EAD directly, instead of using the conventional CCF approach; (2) analyse EAD in the hitherto underresearched area of retail credit cards; (3) apply the idea of EAD mixture models under the GAMLSS framework and compare its performance to a series of benchmark models; (4) identify the factors that significantly impact the mean and dispersion of EAD, giving further insights into the risk drivers of EAD; (5) inspect any differences in the risk drivers depending on whether the account hit the limit prior to default.

The paper is structured as follows. In Section 2, the existing literature on EAD modelling is reviewed. Section 3 explains the data and variables used and Section 4 illustrates how statistical models are constructed. The results are presented and discussed in Section 5. Section 6 concludes.

2. Literature Review

In order to model the EAD of revolving exposures, the Basel II and III Accords have implicitly suggested estimating a Credit Conversion Factor (CCF), which is the proportion of the undrawn amount at the time of estimation (i.e. credit limit minus current balance) that will be drawn by the time of default, i.e.:

$$CCF_{t,\tau} = \frac{\text{EAD}_{t,\tau} - \text{Balance}_t}{\text{Limit}_t - \text{Balance}_t}.$$

Balance_t denotes the amount of money owed by credit card borrowers at the present time (t). Limit_t is the credit limit or maximum amount that the borrowers could draw at t. $\text{EAD}_{t,\tau}$ is simply the balance at the future default time (τ) estimated at the present time (t). Hence,

$$\text{EAD}_{t,\tau} = \text{Balance}_t + CCF_{t,\tau} \times (\text{Limit}_t - \text{Balance}_t).$$

Analysing CCFs (or other EAD proxies that incorporate current balance and limit) is deemed important because current exposure alone does not give a reliable indication of the final balance at default. The reason is that, as obligors are approaching default, they may draw additional money (or, in some cases, pay back part of the balance). Gürtler et al. (2018) found that the most relevant factors affecting CCF are time to default and borrower risk (credit quality). Moreover, CCF values heavily depend on the type of product (corporate or retail), data, and empirical methodology used. In the corporate setting, Gibilaro and Mattarocci (2018) also considered the impact of firms having multiple banking relationships, finding that by considering the exposures as a group rather than individually, one could enhance statistical model fit (in terms of R^2) and reduce the risk of underestimation.

CCF distributions tend to be highly bimodal with a probability mass at zero (when there is no change in balance) and another at one (when borrowers end up drawing the entire limit), while showing a flat distribution in between. This causes difficulties in modelling and predictions produced by a conventional Ordinary Least Squares (OLS) regression model could be poor. Therefore, various techniques and models have been put forward as better alternatives for modelling CCF, e.g., Binary logit and Cumulative logit regression models (Brown, 2011), Beta link generalized linear models (Jacobs, 2010), and Naive Bayesian models and single layer neural networks (Yang and Tkachenko, 2012). Empirical evidence suggests most of these produce better performance than OLS regression.

Even though indirect EAD models based on the CCF are commonly used, several other drawbacks have been identified. For example, when the current drawn balance is already close to the limit to begin with, CCF values can become very large and unstable due to the contracting denominator (or even undefined when balance equals limit). This is not uncommon for accounts that will eventually default. Hence, restrictions must be imposed on CCF models (via truncation or censoring), causing loss of potentially useful information. More well-behaved values could be equally problematic, however. For example, Leow and Crook (2016) pointed out that a positive value of CCF can be observed under two different circumstances: (1) when the current balance is less than both balance at default and current limit (which is a common occurrence for accounts going into default); or (2) when the current balance is greater than both balance at default and current limit. Although these two cases may result in the same positive range of CCF values, their characteristics and implications for EAD risk are totally different. This makes the CCF estimate more difficult to interpret. Furthermore, Taplin et al. (2007) illustrated how predicted values greater than one also create undesirable outcomes. Firstly, they would imply that as the balance increases, EAD (and thus the risk) will decrease, which is counter-intuitive because larger balance should intuitively mean larger exposure. Secondly, when the predicted CCF is greater than one and balance is greater than limit, the estimated EAD would be smaller than both balance and limit, which is unlikely to occur. For regulatory capital requirement purposes, the Basel Accords therefore impose calculated CCF values to be strictly in the [0,1] range. However, in real-life datasets, one can often see a large number of CCF observations that are either negative or exceed one. They could be negative when EAD is less than current balance (i.e. the debtor pays back part of the debt before defaulting), providing that balance is below limit. This more often happens when time to default is large and current credit utilisation is close to one (Moral, 2006). Alternatively, in the empirical dataset analysed by Taplin et al. (2007), 38 percent of all accounts exhibited negative CCF values because they started off with a balance that exceeded the limit (which is contrary to the CCF's core idea of the balance increasing by a fraction of the undrawn amount). Conversely, a sizable proportion of observed CCFs may be greater than one because, in practice, the balance at default time commonly goes beyond the current credit limit, e.g. due to interest and other charges or credit limit increases between t and τ (Tong et al., 2016). Imposing a ceiling on CCF would mean that no EAD estimates could ever exceed the current limit level, which may not reflect reality.

With these obstacles in mind, Luo and Murphy (2020) avoided CCF by implementing other EAD factors, namely EADF ($\text{EAD}_{t,\tau}/\text{Limit}_t$) and AUF ($(\text{EAD}_{t,\tau} - \text{Balance}_t)/\text{Limit}_t$), when estimating EAD in the context of U.S. construction loans. However, these measures might not offer a better alternative. For example, EADF does not incorporate information on the outstanding balance, which the literature consistently finds to be a strongly significant covariate for EAD. Also, Leow and Crook (2016) indicated that, as an account approaches default and balance increases, lenders act differently; some increase the limit level, some reduce it. This leads to a heterogeneity problem in a cross-sectional model. Finally, being ratios as well, these EAD factors share the same weakness as CCF — they become unstable in the case of a small denominator value. Hence, similar restrictions must be imposed for these models (via truncation or censoring).

In light of these drawbacks, alternative approaches have been proposed that involve modelling EAD directly, as a monetary amount (as opposed to ratio). For example, Thackham and Ma (2018) suggested that, for large corporate revolving facilities, banks often actively manage the borrower's limit amount as default time approaches, and that these changes in limit (up or down) have a large impact on EAD. Therefore, they proposed a mixture (two-stage) model, conditioning their EAD target variable on whether the limit is decreased or not. Hon and Bellotti (2016) did not forecast drawn balance at default time (EAD) as such, but instead proposed models to estimate drawn credit card balance at every time step, unconditional on a default event. They argued that, apart from having risk management applications, the prediction of this unconditional balance on revolving credit lines is beneficial because it provides banks an expected profit estimate. Different models were considered, including OLS, two-stage, mixture regression and random effects panel models. The direct EAD model proposed by Tong et al. (2016) uses a zero-adjusted gamma (ZAGA) distribution to capture the EAD distribution observed in a dataset of credit card defaults, grouped per default cohort. They constructed a model in the GAMLSS framework, the predictive performance of which they compared against that of three common CCF models and a utilisation change model. The results confirmed that the direct EAD model is a competitive alternative to these benchmark models. Lastly, another mixture model is proposed by Leow and Crook (2016). Using a portfolio of defaulted credit card accounts and their monthly observations, they analysed outstanding balance. Similarly to Hon and Bellotti (2016), they did so not only at the time of default, but at any time over the entire period up to the default time. In addition, they proposed modelling the probability that account borrowing reaches (or exceeds) the limit level at any time period; under that scenario, they proposed modelling the limit rather than the balance. A discrete-time repeated events survival model and panel models with random effects were applied to estimate the former probability and the conditional balance or limit, respectively, which were shown to provide competitive model fit and predictive accuracy compared to conventional models. As with other such panel models, suitable lags would have to be introduced to make the approach suitable to EAD prediction under Basel, which generally assumes a one-year horizon.

Regardless of the method used to model EAD, common major drivers of EAD according to the literature are commitment limit level, current balance, credit utilisation, credit quality, time to default, and undrawn percentage ($1-(\text{Balance}/\text{Limit})$). In this paper, we use the same variables, supplemented by further behavioural variables derived from monthly account data, as well as a selection of macroeconomic covariates.

Similarly to Tong et al. (2016), our newly proposed direct EAD model is built under the Generalized Additive Models for Location, Scale and Shape (GAMLSS) framework (Stasinopoulos et al., 2017). This framework allows selecting a distribution for the response variable, the parameters of which (location, scale, and shape) can be modelled as a function of explanatory variables, either parametrically or non-parametrically. GAMLSS is much more flexible than the Generalized Linear Model (GLM) or Generalized Additive Model (GAM) frameworks, which are restricted to the exponential family. It potentially allows the fitted distribution to (1) be highly skewed and kurtotic, (2) be discrete, continuous, or mixed discrete-continuous, (3) exhibit heteroscedasticity, whereby the value of scale and shape parameters varies across covariate levels. This is important for observed EAD data as it typically exhibits several of these features. Moreover, the ability to model the dispersion of EAD as a function of explanatory variables can be useful from a risk management perspective; where the estimated EAD dispersion is large, we could thus make the point estimate more conservative in order to deal with the greater uncertainty. Motivated by the empirical results reported by Leow and Crook (2016), we further extend the approach by considering that, as accounts move towards default, the balance could either hit the limit or not. This breaks the EAD model into two mixture components, which could have different EAD levels and risk drivers. Although considering similar scenarios, our approach differs from that taken by Leow and Crook (2016) in a number of ways. First, rather than treating balance as panel data, we apply the default cohort approach in EAD modelling and group defaults according to 12-month calendar periods, as this facilitates producing estimates that are conditional on default and matches the prediction horizon used for Basel. Second, using the GAMLSS framework for all model parts offers a wider range of distributions and, importantly, allows introducing non-linearity. Third, considering that the balance can further vary over time and may exceed the (prior) credit limit, we do not fix the EAD to the credit limit value conditional on a max-out event, but allow its distribution to be explicitly modelled in this mixture component as well, thus giving further insights into specific risk drivers of EAD for this subgroup.

Note that our proposed model is an account-level one; in other words, it is the result of taking a bottom-up approach. More generally, the underlying parameters in credit risk modelling can be estimated in two different ways: top-down or bottom-up (BCBS, 1999). The former approach aggregates data with similar risk profiles, e.g. with regard to credit rating and tenure, and groups them into homogeneous pools, for which well-calibrated credit risk parameter estimates are then provided. This method is typically applied to consumer, credit card or other retail portfolios, due to their volume. For example, Witzany (2011) showed how EAD could be estimated at the aggregated pool level by the top-down approach. On the other hand, the bottom-up approach measures credit risk at an individual (loan or account) level, considering information on the entire set of (inhomogeneous) loans. This approach is often adopted for corporate exposures and capital market instruments. In the consumer credit risk literature, both of these approaches are well known and have each been employed; however, one does not rule out the other. For example, the bottom-up approach could aid the design of top-down models as it allows loans to be classified into pools using individual loan data, whilst the pool-level risk parameter could eventually be estimated from the aggregated data. Since all of the individual card defaults are used to construct the EAD models in this paper, our method would be primarily classified under the bottom-up approaches. Examples of other studies that, similarly to us, utilise a bottom-up method for retail credit card modelling are Tong et al. (2016), Hon and Bellotti (2016) and Leow and Crook (2016).

The use of regression models, in line with a bottom-up approach, offers a number of benefits. First, they provide insight into any account-level or economic risk drivers of EAD. Second, they are able to produce forward-looking estimates, and can thus capture the effects of any changes in the underlying risk drivers; from this, one can project how total portfolio risk changes if the portfolio composition (due to changing customer characteristics or account behaviour) and/or the economy changes. Third, when building account-level models, positive correlation between default risk and EAD can be taken into account in the respective models through a subset of shared covariates such as credit utilisation and economic variables. If, instead, this risk dependence were to be neglected, portfolio risk could be underestimated since it is the underlying force that significantly increases tail losses (see, for example, Kaposty et al. (2017) and Kupiec (2008)).

3. Data and variables

The original dataset provides monthly account-level data on the consumer credit cards of a large Asian bank from January 2002 to May 2007. We define EAD as the outstanding balance at default time, taking the amount owed by the borrower excluding any subsequent interests and additional fees; any debt incurred after default will not be included in the EAD calculation. We say that an account goes into the default state when a borrower either: (1) misses or could not make the minimum repayment amount required by banks for three months or more; (2) is declared bankrupt; or (3) is declared charged-off, i.e. expected to be unable to return the owed money back to the bank. In keeping with the standard practice in EAD modelling, we extract data from the defaulted accounts only, as the estimation is conditional on default and the balances of defaulted and non-defaulted accounts are expected to behave differently.

We also add macroeconomic variables to the dataset because individual customers' borrowing levels could further vary under different economic scenarios. Also, this may help our model be more time-stable and allows us to assess the impact on EAD of downturn scenarios, thus providing a suitable framework for stress testing required by banks applying the A-IRB approach (Kaposty et al., 2017).

We apply the standard yearly cohort method (Moral, 2006) to prepare the data for analysis and set the reference month where the estimation takes place on 1st November of every year. The values of behavioural and macroeconomic covariates are then collected a month prior to the reference month, namely in October,

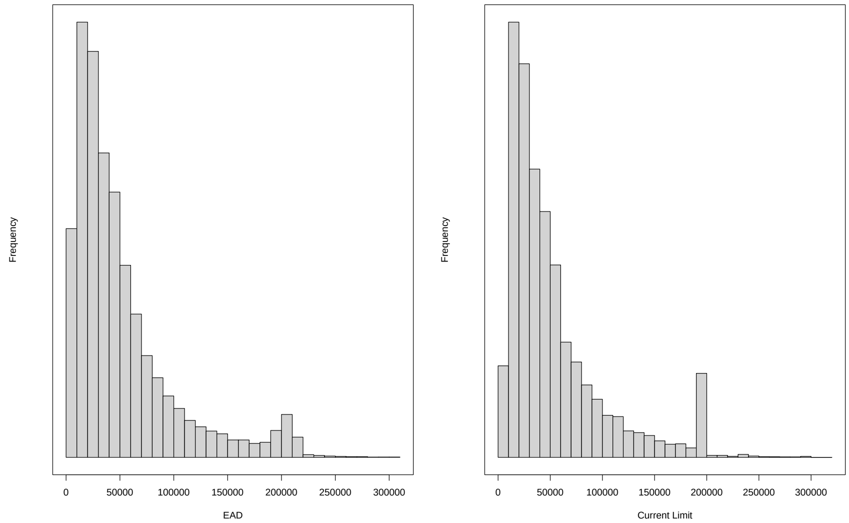

Figure 1: Histograms (with the scale of the y-axis omitted) of: observed exposure at default (left); observed current limit (right).

whereas the response, EAD, is recorded at the occurring default time. Accounts that lack sufficient monthly records to calculate the explanatory variables are omitted.

Further removing a small number of missing value cases (177 observations), we are left with more than 70,000 defaulted accounts. This dataset is then separated into three groups: training (60%), validation (20%), and test dataset (20%). Figure 1 shows the empirical EAD distribution, which exhibits right-skewness, is heavy-tailed, and has a small bump at 200,000 (local currency), which is a likely consequence of the bank operating a maximum limit.

Table 1 lists the set of candidate explanatory variables extracted from the data, which have previously shown correlation with EAD according to the literature or can be reasonably expected to significantly impact EAD. Four macroeconomic variables are considered: unemployment rate, interest rate, GDP, and CPI.

4. Statistical models

The following subsections outline our newly proposed model, GAMLSS.Mix, and three benchmark models, GAMLSS, OLS.Mix and OLS.

4.1. GAMLSS.Mix

In our new model, we propose to estimate EAD conditionally on two mutually exclusive scenarios that may occur in the run-up to default. Denote the EAD of account i as $\text{EAD}_{t,\tau}(i) = \text{EAD}_i$. Note that reference time t and default time τ are omitted from here on for the sake of simplicity. We define a binary variable, S_i, to denote the occurrence of a "max-out" event as:

$$S_i = \begin{cases} 1 & \text{if the balance hit the limit, at least once, at any point during the outcome window;} \\ 0 & \text{otherwise,} \end{cases}$$

where the outcome window is the period between reference and default time. Applying the law of conditional expectation, the expected value of EAD_i is then given by:

$$E(\text{EAD}_i) \;=\; [P(S_i = 1) \times E(\text{EAD}_i|S_i = 1)] + [P(S_i = 0) \times E(\text{EAD}_i|S_i = 0)]\,. \tag{1}$$

Therefore, three model parts must be fitted, all of which conditional on default: first, a model for the probability that the balance will hit the limit over the observation period, $P(S_i = 1)$; second, a model to estimate EAD conditional on the balance hitting the limit, $E(\text{EAD}_i|S_i = 1)$; third, a model to estimate EAD conditional on no such max-out event occurring, $E(\text{EAD}_i|S_i = 0)$.

We will refer to this newly proposed mixture model as "GAMLSS.Mix", as it will use the GAMLSS framework to fit each of these model parts. For each such component, we use a separate validation set (setting aside 20% of the data) to make model selection decisions such as variable selection. Subsequently, the final model (whose partial effect plots will be shown) is trained after merging training and validation data.

Variable	Notation	Explanation
Age of account	age	Months since account has been opened.
Limit	l	Credit limit, i.e. maximum amount that could be drawn from card.
Balance	b	Current amount drawn.
Behavioural score	bsco	Internal score capturing current credit quality of account.
Months in arrears past 9 months	no.arr9	The number of months an account has been in arrears over the nine months prior to the reference time. A borrower is considered in arrears when they pay less than their monthly minimum payment.
Months in arrears past 3 months	no.arr3	
Limit increase past 9 months	limin9	Dummy variable indicating whether the limit has been increased over the past nine months (Y/N).
Limit increase past 3 months	limin3	
Absolute balance change past 9 months	abs.ch.b9	
Absolute balance change past 3 months	abs.ch.b3	
Average paid percentage past 9 months	paid.per9	Paid percentage is the percentage of last month's balance paid by the borrower, i.e. Paid Amount/Balance.
Average paid percentage past 3 months	paid.per3	
In arrears past 9 months	arr9	Dummy variable indicating whether the account has been in arrears at least once over the past nine months (Y/N).
In arrears past 3 months	arr3	
Credit utilisation	cu	Percentage of the limit drawn by borrower, i.e. Balance/Limit.
Full payment percentage	full.pay.per	Percentage of account's months on book in which borrower has paid balance in full, i.e. number of full payments / age of account.
Behavioural score special code	bscocat	Dummy variable indicating whether behavioural score recorded a "special" case.
(Non-)negative balance	bcat	Dummy variable indicating whether balance was negative (and thus capped).
Time to default	ttd	Duration in months from reference time to default time.
Unemployment rate	unem	Macroeconomic variable measured at reference time.
Interest Rate	int	The best lending rate benchmarked by the relevant central bank.
Gross domestic product	gdp	
Consumer price index	cpi	

Table 1: List of available explanatory variables. Note that, since the behavioural scores of some accounts do not have a regular value (such as 680, 720, etc.) but codes representing "special" cases (e.g., "the account is too new to score"), we replace such special codes by the (training) mean of the regular behavioural scores and flag this up with the help of a dummy indicator (bscocat). Likewise, negative credit card balances, which may e.g. occur when a borrower uses a credit card to purchase a product and decides later to return it, are capped at zero, and another dummy variable (bcat) is added to distinguish between negative and true zero balances.

4.1.1. Probability of max-out event

To estimate $P(S_i = 1)$, we model the binary response variable as a non-parametric function of the explanatory variables. More specifically, letting $p_i = P(S_i = 1)$, the max-out event probability is modelled as follows:

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha_1 Y_{i,t} + \alpha_2 Z_t + \text{non-parametric terms}, \tag{2}$$

where α_1 and α_2 are unknown vectors of parameters to be estimated, and $Y_{i,t}$ and Z_t are (account-level) behavioural and macroeconomic covariate vectors, respectively. The parametric coefficients α_1 and α_2 and non-parametric smoothing terms are fitted by performing the Rigby and Stasinopoulos (RS) algorithm based on penalised (maximum) likelihood (Stasinopoulos et al., 2017), into which the following likelihood function, L, is substituted:

$$L = \prod_{i=1}^{n} p_i^{y_i} \times (1-p_i)^{1-y_i}, \tag{3}$$

where $y_i = 1$ for an observation i whose balance hit the limit, and zero otherwise. Penalised B-splines (Eilers and Marx, 1996) are chosen to fit the non-parametric terms in Equation 2 because they enable smoothing parameter selection to be performed automatically by minimising the Akaike Information Criterion, AIC $= -2L^p + 2n$, where L^p is the penalised likelihood and n is the number of parameters in the model.

We build three candidate models for p_i by considering three different variable selection strategies — either including all explanatory variables or using two alternative stepwise methods (alternating forward and backward selection at each step) based on the AIC and BIC criteria (5 percent α-level). The "gamlss" package (Stasinopoulos et al., 2017) in R (R Core Team, 2020) is used to fit these three models to all training examples of defaults. Based on their performance on the validation set, one of the three candidate models is then selected, following assessments of the Pearson goodness-of-fit statistic from the Hosmer-Lemeshow test (predictive accuracy), Area Under the Receiver Operating Characteristic curve (AUROC) (discrimination power) and residual plots (model adequacy). Where these metrics suggest different candidate models, one is chosen at the modeller's discretion. Note that the residuals used in GAMLSS are normalised quantile residuals which are expected to follow a standard normal distribution regardless of the distribution of the response variable, provided that the model is correctly specified.

4.1.2. Conditional EAD models

To produce EAD estimates that are conditional on either of the two credit balance scenarios, we further partition the training data into two subsets. The first subset consists of the credit card accounts whose balance hit the limit in any of the months during the outcome window; the second subset consists of the accounts that did not. We then proceed by fitting two separate models to these subsets.

In either of these scenarios, one can further distinguish between zero and non-zero EAD values. Zero values may potentially occur because of several special cases or technical default examples, such as charge-offs connected to other accounts, the observations being rounded or truncated to zero, customers moving their outstanding balance to other accounts, or payment delays. As they could have different explanatory drivers, we treat zero values separately from non-zero EAD values by including the probability of zero EAD into the models.

Figure 2 shows the empirical distribution of non-zero EADs for both subsets of accounts, confirming that accounts that hit their limit tend to have larger EAD values. Their shape also suggests a positively skewed distribution such as Gamma, Inverse Gaussian, or Log Normal distribution. For each of these candidate distributions, we evaluated the AIC/BIC and MAE/RMSE criteria for a full model (i.e. with all explanatory variables). Based on this, as in Tong et al. (2016), the Gamma distribution was found to give the best results.

Hence, in order to model $E(\text{EAD}_i|S_i = 0)$, we assume that EAD_i follows a mixed discrete-continuous Zero-Adjusted Gamma (ZAGA) distribution, shown in Equation 4.

$$f(\text{EAD}_i|S_i = 0) = \begin{cases} \nu_i & \text{if } (\text{EAD}_i|S_i = 0) = 0, \\ (1-\nu_i)\,\text{Gamma}(\text{EAD}_i|\mu_i, \sigma_i, S_i = 0) & \text{if } (\text{EAD}_i|S_i = 0) > 0, \end{cases} \tag{4}$$

for $0 \leq \text{EAD}_i < \infty$, where $0 < \nu_i < 1$, $\mu_i > 0$, $\sigma_i > 0$, and

$$\text{Gamma}(y|\mu,\sigma) = \frac{1}{(\sigma^2\mu)^{1/\sigma^2}} \frac{y^{\left(\frac{1}{\sigma^2}-1\right)} e^{-y/(\sigma^2\mu)}}{\Gamma(1/\sigma^2)}.$$

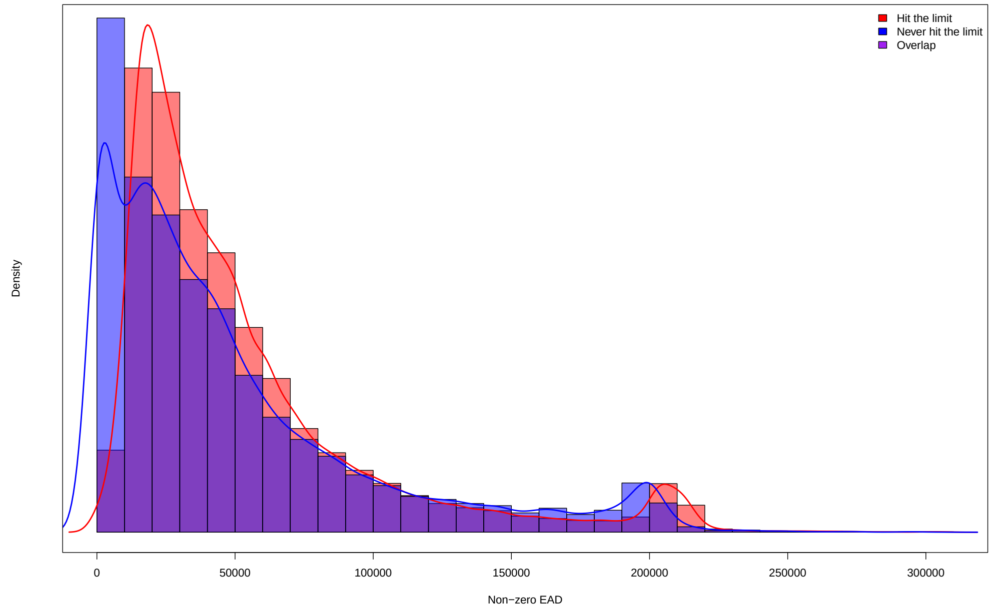

Figure 2: Empirical distribution of non-zero EADs (with the scale of the y-axis omitted); red: histogram for the accounts whose balance hit the limit, blue: histogram for the accounts that never hit the limit; purple: overlapping area.

Note that the mean and variance of Gamma$(y|\mu,\sigma)$ are μ and $\sigma^2\mu^2$, respectively. Hence,

$$\begin{aligned} E(\text{EAD}_i|S_i=0) &= (1-\nu_i)\mu_i, \\ Var(\text{EAD}_i|S_i=0) &= (1-\nu_i)\mu_i^2(\sigma_i^2+\nu_i). \end{aligned} \tag{5}$$

There are thus three parameters in the ZAGA distribution: the mean (μ) and dispersion (σ) of non-zero EAD, and the probability of zero EAD (ν). Allowing the relationship between μ and its explanatory variables to be non-linear, we again model it through non-parametric smoothing terms. Since the main focus is on μ, we restrict the relationships of σ and ν with their respective sets of explanatory variables to be parametrically linear. This makes the model less computationally expensive and easier to implement in practice. The parameters μ, σ, and ν can thus be estimated through the following link functions:

$$\log(\mu_i) = \gamma_1^{\mu} Y_{i,t}^{\mu} + \gamma_2^{\mu} Z_t^{\mu} + \text{non-parametric terms};$$

$$\log(\sigma_i) = \gamma_1^{\sigma} Y_{i,t}^{\sigma} + \gamma_2^{\sigma} Z_t^{\sigma}; \quad \text{logit}(\nu_i) = \gamma_1^{\nu} Y_{i,t}^{\nu} + \gamma_2^{\nu} Z_t^{\nu},$$

where γ_1 and γ_2 are unknown vectors of parameters to be estimated. We apply a log and logit link function, respectively, in order to assure that the range of μ and σ parameters are greater than zero and the range of ν parameter is between zero and one. The likelihood function, L, used in the penalised maximum likelihood estimation is:

$$L = \prod_{i=1}^{n} f(\text{EAD}_i) = \prod_{\text{EAD}_i=0} \nu_i \prod_{\text{EAD}_i>0} (1-\nu_i) \times \text{Gamma}(\text{EAD}_i|\mu_i,\sigma_i). \tag{6}$$

Five variable selection techniques are applied to create five submodels: using all variables; using stepwise variable selection for μ, σ and ν separately, with either AIC or BIC as the model selection criterion; using stepwise with AIC/BIC by running the parameters together (cf. stepGAICAll.A() function in Stasinopoulos et al. (2017)). The criteria used to select one of the five resulting submodels are Pearson correlation (discrimination performance), MAE, Normalised MAE, RMSE, Normalised RMSE (predictive accuracy) and residual plots (model adequacy), each of which is again evaluated on the validation set. Normalised versions of the error metrics are produced where MAE and RMSE are calculated for EAD/Current Limit, instead of EAD, in order to investigate the performance of the model if the percentage of current limit (not EAD itself) at default time is of interest.

The procedure of modelling EAD for the second subset of accounts that hit their limit, $E(\text{EAD}_i|S_i = 1)$, is similar to its counterpart scenario. The Gamma distribution is again selected for fitting the non-zero EAD response. Note that the best model variations for all three model components can be found in Table 2.

Model	Probability max-out	EAD no max-out	EAD max-out
GAMLSS.Mix	Full variables	Stepwise AIC, run separately for each parameter	Stepwise BIC, run separately for each parameter
GAMLSS	Stepwise, with BIC, run for all model parameters together		
OLS.Mix	Stepwise with BIC	Full variables; LASSO	Stepwise with AIC
OLS	Full variables; LASSO		

Table 2: Best submodels for the newly proposed and benchmark models.

4.2. Benchmark models

In order to evaluate the effectiveness of our proposed model, we build another three benchmark models against which we compare its predictive performance. Firstly, "GAMLSS" is the EAD model under the GAMLSS framework applied to all defaulted accounts, without applying the mixture idea. Secondly, "OLS.Mix" adds the mixture idea to the OLS framework, applying standard OLS regression for the mixture components and logistic regression for modelling the max-out event probability. Thirdly, "OLS" fits a standard OLS regression model to all defaulted accounts. To perform variable selection for OLS and OLS.Mix, we try three methods: the Least Absolute Shrinkage and Selection Operator (LASSO), a stepwise algorithm, and fitting a model with the full set of variables. As before, we use a validation dataset to find the best (sub)model candidates for each benchmark approach (see Table 2).

5. Results and discussion

In this section, we present the results of our newly proposed model and the performance comparisons with the benchmark models. In addition, we will inspect the significant relationships between explanatory variables and response parameters.

5.1. Discrimination and predictive performance

The performance measurements for all models, evaluated using ten-fold cross validation, are shown in Table 3. This table contains the following metrics: Pearson correlation (discrimination performance); MAE, Normalised MAE (see section 4.1.2), RMSE, Normalised RMSE (predictive accuracy); and 0.9 quantile loss (QL-90). The α quantile loss function is defined as $\sum_{i;y_i<\hat{y}_i}(\alpha - 1)\cdot(y_i - \hat{y}_i) + \sum_{i;y_i\geq\hat{y}_i}\alpha\cdot(y_i - \hat{y}_i)$, where y_i and $\hat{y}_i$ are true and predicted EAD values, respectively. Its basic idea is to give different penalties to a misestimation based on the selected quantile. The 0.9 quantile loss penalises underestimation more heavily, and hence, is a good measure for assessing the conservativeness of a risk estimate such as EAD.

As time to default is unknown a priori, Table 3 presents two different sets of results: one using the actual values of time to default (to enable comparison with other papers that included this variable and as it is likely to affect dispersion); and one where they were estimated by applying a simplified version of the PD-weighted approach by Witzany (2011), in which, for each month t ($t = 1, \ldots, 12$) of each default cohort, we observe the empirical proportion of training set defaults, $\text{PD}(t_i)$ and, from those, derive the following point prediction for EAD of each account:

$$\text{EAD} = \sum_{t_i=1}^{12}\left[\text{PD}(t_i) \times \text{EAD}(t_i)\right], \tag{7}$$

where $\text{EAD}(t_i)$ is the EAD estimate when t_i months is substituted instead of the actual time to default. The latter approach is used to verify to what extent the former performance results remain robust if the model

Model	Correlation	RMSE	MAE	Norm.RMSE	Norm.MAE	QL-90
GAMLSS.Mix	0.937 (0.004)	16927 (385)	7881 (137)	0.265 (0.005)	0.147 (0.002)	4125 (153)
GAMLSS	0.908 (0.041)	20489 (4903)	8718 (268)	0.292 (0.010)	0.160 (0.002)	4457 (161)
OLS.Mix	0.935 (0.004)	17152 (435)	8751 (161)	0.304 (0.011)	0.187 (0.003)	4365 (159)
OLS	0.930 (0.004)	17810 (397)	9758 (162)	0.335 (0.009)	0.220 (0.004)	4879 (155)

(a) Performance measurements, using actual values of time to default.

Model	Correlation	RMSE	MAE	Norm.RMSE	Norm.MAE	QL-90
GAMLSS.Mix	0.933 (0.004)	17458 (381)	8136 (140)	0.273 (0.006)	0.151 (0.002)	4354 (153)
GAMLSS	0.907 (0.041)	20591 (4882)	8757 (268)	0.293 (0.010)	0.161 (0.002)	4471 (158)
OLS.Mix	0.932 (0.004)	17574 (434)	8845 (159)	0.298 (0.009)	0.183 (0.003)	4532 (154)
OLS	0.929 (0.004)	17945 (394)	9500 (147)	0.315 (0.007)	0.203 (0.003)	4750 (146)

(b) Performance measurements, using PD-weighted approach.

Table 3: Ten-fold cross validation performance measurements with standard errors inside parentheses.

is applied not for explanatory (using real values of time to default) but for prediction purposes (using the estimated values).

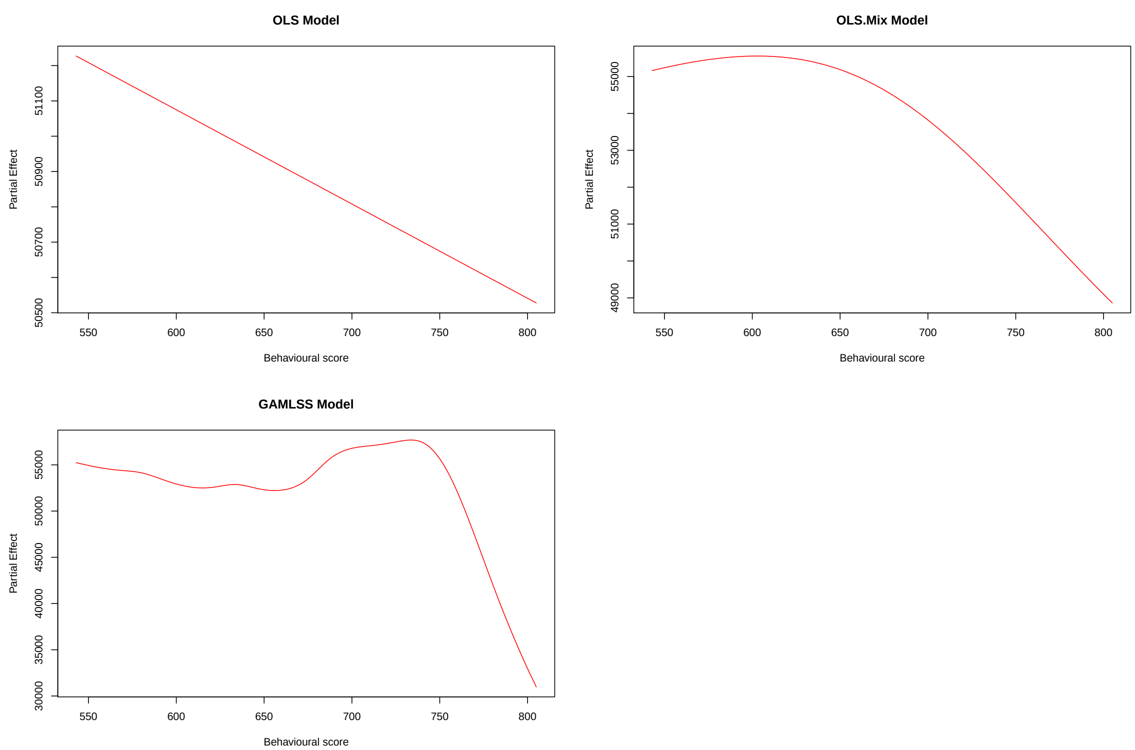

Figure 3: Partial effect plots of behavioural score vs. estimated EAD, for OLS, OLS.Mix and GAMLSS models.

Examining the results, we can see that, when it comes to Pearson correlation, there is little to separate the different models, indicating that even the simplest model (OLS) can already discriminate well between high and low EAD risk. However, with regards to all other measures, there are pronounced differences between the various approaches. Firstly, with the exception of RMSE and QL-90 for GAMLSS, the two approaches that apply the GAMLSS framework (GAMLSS and GAMLSS.Mix) outperform those using standard OLS regression (OLS and OLS.Mix), showing that its features are better capable of handling the EAD distribution and its relation to the risk drivers (e.g. any non-linearity). Secondly, when we introduce the mixture concept into the OLS framework (OLS.Mix vs. OLS), all of the predictive accuracy measures improve as well. This is in agreement with the results reported by Leow and Crook (2016), who also found that adding the mixture component to their linear models improved performance. We suggest, as a partial reason for this performance gain, that conditioning on the occurrence of a max-out event has the beneficial effect of introducing some non-linearity into the functional relationships between explanatory variables and EAD. This is illustrated by the partial effect plots for the behavioural score variable in Figure 3, showing us how OLS.Mix is able to approximate the non-linear relationship between behavioural score and EAD using a concave function. Thirdly, and perhaps most importantly, the newly proposed model, GAMLSS.Mix, consistently outperforms all benchmark models across all predictive performance criteria (cf. RMSE, MAE, Norm.RMSE, Norm.MAE), whilst being more conservative in terms of the prediction errors it makes (cf. QL-90). This shows that, as hypothesised, there is indeed added value in combining both modelling elements.

When comparing the predictive performance without prior knowledge of time to default (see Table 3b) against that of the explanatory model application (i.e. with knowledge of time to default, see Table 3a), we see a small drop in performance, as to be expected, but importantly, the performance ranking for all models remains almost similar and the proposed GAMLSS.Mix model still has the best predictive power. This suggests that our findings are robust regardless of the chosen treatment of this explanatory variable.

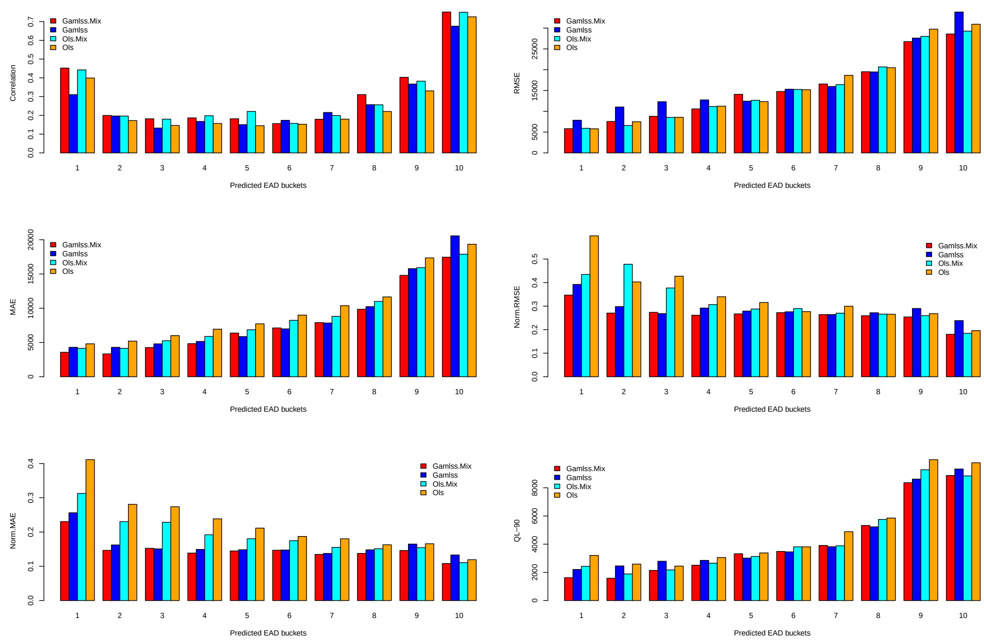

Figure 4: Performance measurements assessed on different buckets of the test dataset, ordered from lowest (1) to highest (10) predicted EAD; results using actual values of time to default.

Since the proposed model is designed using the bottom-up approach, it is interesting to verify how effective it is in producing predictions in different risk buckets. Hence, for each model, we sort the accounts in the test set into ten different buckets according to the value of predicted EAD (the first bucket containing those accounts with the lowest 10% predicted EAD, while the tenth bucket holds the highest 10% ones). Then, we evaluate each model's performance in each of these buckets (see Figure 4 when actual time to default is used and Figure 5 when the simplified PD-weighted approach is used). The results show that, most of the time, the GAMLSS.Mix model provides the best performance across the different measures. Therefore, the proposed bottom-up model also performs well when applied to the accounts in different pools. In addition, we provide calibration plots showing how close the bucket-level EAD estimates produced by each method are to the actual (mean) EAD in each bucket (see Figure 9 and Figure 10 in appendix). All (bottom-up) models are shown to be well calibrated at each (top-down) EAD pool.

5.2. Risk Drivers of GAMLSS.Mix model components

Unlike with a linear regression, the non-parametric smooth functions fitted by GAMLSS.Mix cannot be explained in a simple mathematical form; that is, we cannot gauge the impact of an explanatory variable

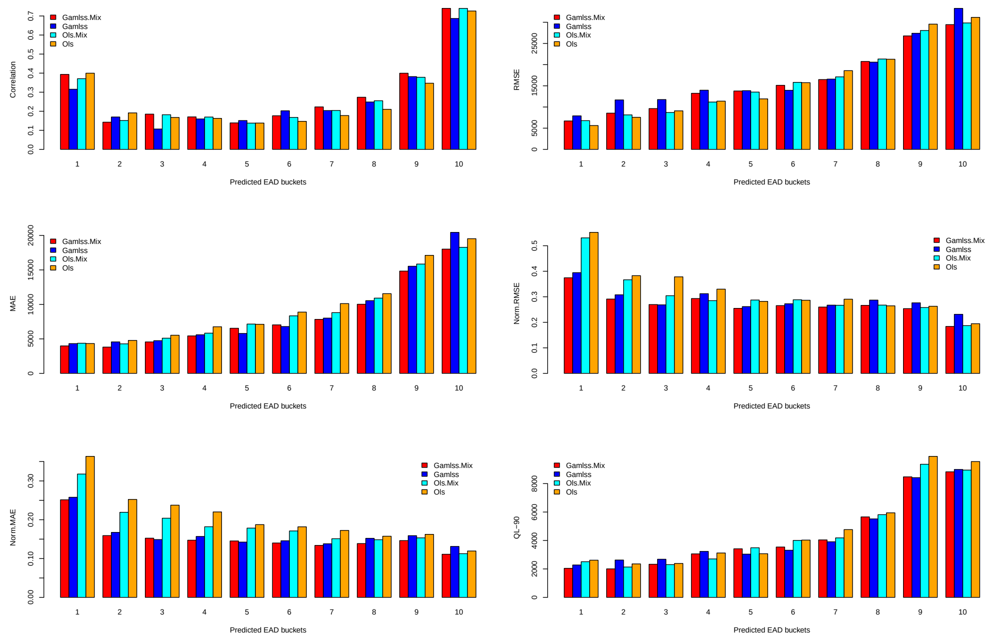

Figure 5: Performance measurements assessed on different buckets of the test dataset, ordered from lowest (1) to highest (10) predicted EAD; results using the simplified PD-weighted approach.

on the response variable by just looking at its estimated coefficient. However, we can display each effect visually with the help of partial effect plots. These depict how one specific explanatory variable influences the response assuming that the other covariates are fixed.

Figure 6 displays partial effect plots on a logit scale for the max-out event probability, $P(S_i = 1)$, of GAMLSS.Mix. The shaded areas indicate the precision of the estimates using 95% confidence intervals. In the bottom-right panel, we observe that higher credit utilisation (measured at reference time) makes it more likely that the customer will max out their card in the run-up to default, especially when utilisation already exceeds one prior to the outcome period (the latter makes the event almost inevitable). Similarly in line with expectations, longer time to default (see bottom-left panel) is associated with a higher probability of the balance hitting the limit. Starting balance (top-left) and credit limit (top-right) tend to have a positive and negative effect on the probability of a max-out event, respectively, which is again intuitive since customers with higher balance and lower limit are closer to maxing out their card.

Figure 7 presents the partial effect plots, on a log scale, for the μ parameter (non-zero EAD mean) of GAMLSS.Mix, for the subset of accounts whose balance never hit the limit (hence, conditional on $S_i = 0$); Figure 8 does so for the other subgroup ($S_i = 1$). In both figures, we see that higher credit limit level is strongly linked to larger EAD. This is again perfectly intuitive as customers with a higher limit are allowed to borrow more. Note that the waviness and widening confidence band near the upper-end of the variable range suggest some undersmoothing linked to the relatively small number of accounts with a limit above 200,000.

Similarly, EAD is also related to the current level of credit utilisation (Figure 7, top-left plot) or to balance (Figure 8, left plot), higher values implying larger balance at default. Interestingly, more variables appear in Figure 7, suggesting that these only help to better predict accounts who stay clear of the limit. In other words, a more complex model is needed for this mixture component than for the other. For example, in their higher value range, behavioural score and full payment percentage have a negative effect on the EAD of those accounts; hence, provided that they did not hit the limit, high credit-quality borrowers who most of the time pay back their balance in full tend to have a lower balance if they do default. Two novel insights were encountered as well. Firstly, the partial effect plot for average paid percentage over the past three months (see top-middle panel of Figure 7) suggests that those borrowers who previously repaid a higher (partial) proportion of their balance could still end up with a higher EAD. Secondly, customers with a

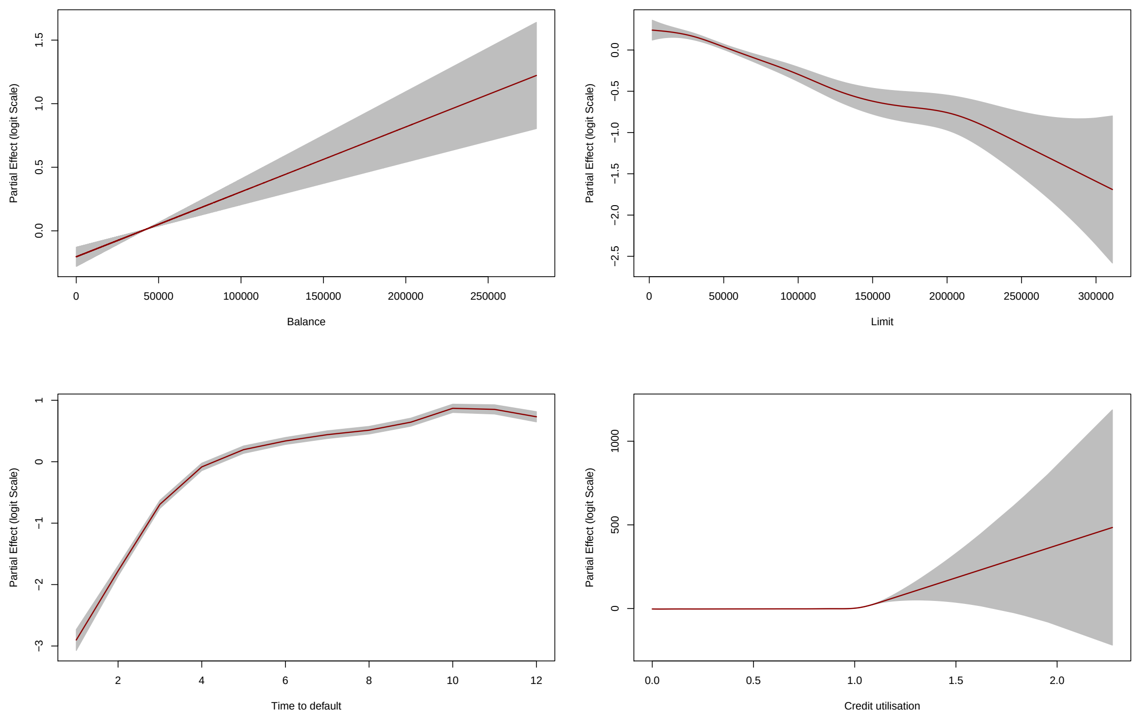

Figure 6: Partial effect plots on logit scale for max-out event risk in the GAMLSS.Mix model.

negative current card balance (who are thus owed money by the bank) may have higher EAD risk than those with zero balance (top-right). One potential explanation may lie in that both could be seen as indicative of greater card activity. Another may be that, as those values are more often associated with customers who are less likely to default, there may be hidden risks that drive them to heavily draw down before default eventually occurs; this would concur with Barakova and Parthasarathy (2013) who reported that higher EAD can be associated with defaults that are hard to anticipate. Note that, for brevity and as they had a lesser impact (based on a likelihood ratio test), macroeconomic covariates and the other behavioural variables are omitted from the figure (results available on request).

To facilitate further comparison between the different models and the effects they captured, Table 4 summarises which explanatory variables are shown to have a strong impact on (non-zero) EAD mean (μ) in the two GAMLSS.Mix component models and the GAMLSS benchmark model and whether that impact is (mostly) positive or negative. Likewise, it also contains the same information for the σ (dispersion) and ν parameters. For brevity, we omit further discussion of the last parameter, ν.

Turning to the second parameter, dispersion, we can see in Table 4 that the higher is the level of credit utilisation and/or current balance, the lower is the dispersion — in other words, the more predictable the EAD. In contrast, the farther away from default time (both scenarios) or the larger the limit (non-max-out scenario only), the larger the dispersion; i.e. there is more time and scope for the balance to change and thus become less predictable. These four effects all appear to be intuitive. Interestingly, as for the EAD mean earlier, the list of important factors is again longer for the first mixture component (i.e. for the accounts with no recorded max-out event). There, age of account (i.e. time on book), the average of paid percentage over three months, and number of months in arrears are also among the variables that are shown to affect dispersion. Specifically, the longer the account has been on the books, the more predictable is EAD, whereas higher values for the other two variables (which could indicate greater monthly variation in balance) tend to imply greater variance. Also, special behavioural scores and negative current balances imply special cases under which the EAD prediction for those accounts becomes more uncertain as well. As they are all meaningful effects, there appears to be added value in explicitly modelling the dispersion parameter (rather

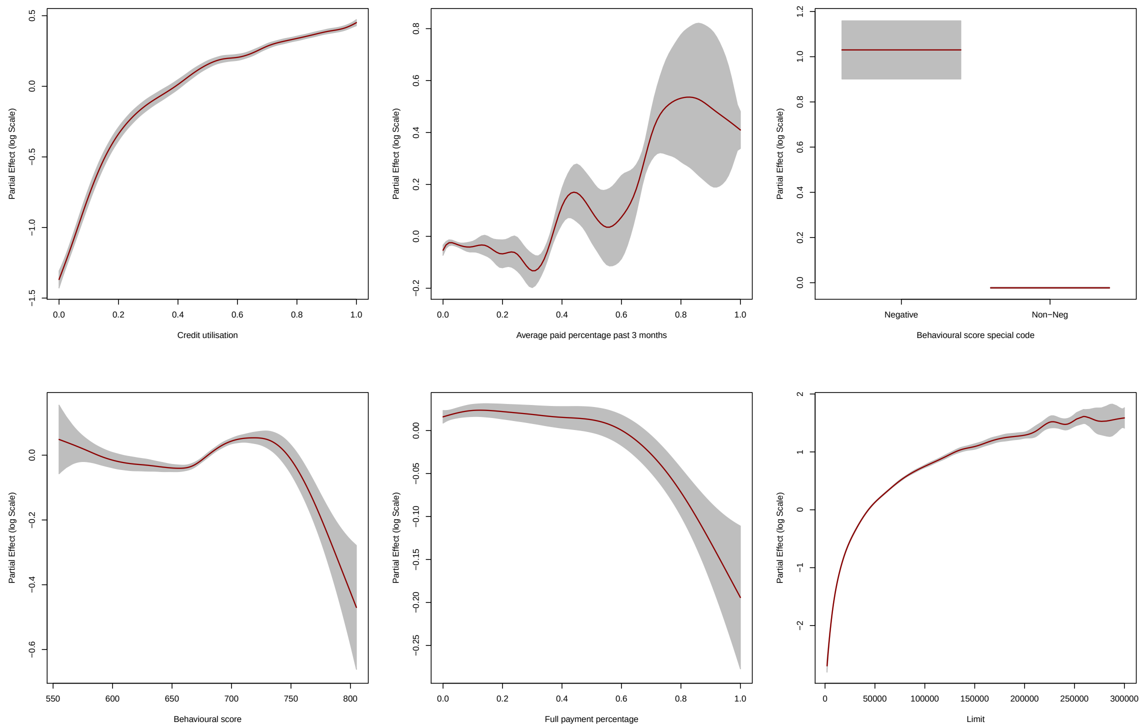

Figure 7: Partial effect plots on log scale for the mean (μ) parameter of the accounts whose balance never hit the limit in the GAMLSS.Mix model.

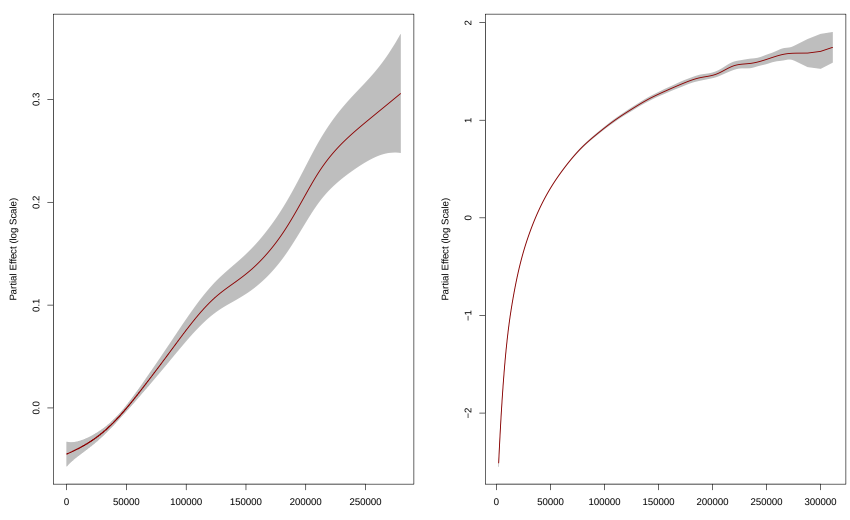

Figure 8: Partial effect plots on log scale for the mean (μ) parameter of the accounts whose balance hit the limit in the GAMLSS.Mix model.

than assuming homoscedasticity).

Variable	Mean			Dispersion			Prob. of zero-EAD		
	EAD	**EADn**	**EADt**	**EAD**	**EADn**	**EADt**	**EAD**	**EADn**	**EADt**
age					−			+	
l	+	+	+	+	+			+	
b	+		+	−	−	−	−		
bsco	−	−		−			+	+	
no.arr9					+				
no.arr3								−	
limin9									
limin3									
abs.ch.b9						+	−	−	
abs.ch.b3				+					
paid.per9									
paid.per3	+	+			+			+	+
arr9									
arr3									
cu		+			−			−	
full.pay.per		−						+	
bscocat (special)				+	+				
bcat (negative)	+	+		+	+				
ttd				+	+	+		−	
unem									
int								−	
gdp									
cpi									

Table 4: A set of strongly significant predictors for the EAD parameters of: the GAMLSS benchmark model (EAD); GAMLSS.Mix no max-out (EADn); and GAMLSS.Mix max-out (EADt).

6. Conclusions and future research

Exposure At Default (EAD) is one of the key parameters used to calculate the regulatory capital requirements under the Advanced Internal Rating Based (A-IRB) approach. To estimate EAD, Credit Conversion Factor (CCF) models were implicitly suggested by the Basel Accords and have been studied in the literature, but several drawbacks of such models can prove problematic. In this paper, we therefore mainly focus on estimating EAD via a direct model rather than applying CCF or other related factors.

Our newly proposed model combines two ideas formerly put forward in the literature. First, it is built under the GAMLSS framework which produces a much more flexible fitted distribution than the GLM and GAM frameworks. Second, as the level of EAD as well as the risk drivers of its mean and dispersion parameters could significantly differ depending on whether the account hit the credit limit at any point in the run-up to default, we extend our solution to a mixture model conditioning on these two possible scenarios. This new model, as well as several benchmark models, are empirically validated using a large dataset of credit card defaults not previously used in the EAD literature.

By distinguishing between these two scenarios, we indeed found differences in preferred risk drivers for the EAD model parameters. For example, current balance was picked over several other potential drivers for (positive) EAD mean when a max-out event occurs, but not in the opposite scenario, whereas current limit level was identified as being strongly linked to dispersion only under the non-max-out scenario. Moreover, the number of factors is larger for borrowers who did not max out their cards, suggesting that this subgroup benefits from a more complex model. Overall, only behavioural variables appear to have a significant impact in our EAD models; despite the data containing defaults from a recessionary period, the macroeconomic covariates show little added predictive power over those account-level variables. Current limit is the strongest variable that affects the mean of non-zero EAD. To manage model uncertainty, one should focus on the current level of drawn balance amount and (estimated) time to default as their values greatly impact EAD dispersion.

Our results show a clear performance benefit of applying GAMLSS over the OLS framework, confirming, consistently with what Tong et al. (2016) reported for another dataset, that there are indeed predictive accuracy gains in EAD modelling from including non-linear effects and targeting not only the EAD mean but also dispersion. Similarly, when the mixture concept is introduced into the OLS framework, all predictive accuracy measures improve as well. A new explanation we put forward for the latter is that, by implementing the mixture idea, we allow some non-linear effects to emerge from the combination of different linear models, thus capturing more complex relationships between EAD and its covariates and producing better predictions. Most of all though, we find that combining the mixture component and the GAMLSS approach results in another predictive performance boost, as our newly proposed model, GAMLSS.Mix, outperforms the three benchmark models on all criteria.

In terms of potential practical benefits, a more accurate EAD model, such as that proposed, can lead to more accurate loss estimation, which allows banks to adjust the capital they require accordingly. Moreover, the non-linear predictor effects, shown in the partial effect plots, reveal the impact of each behavioural variable on different risk aspects. This can provide the bank with useful insights to design an early warning system. More specifically, the insights from the "max-out" model allow the bank to identify those borrowers who are most at risk of maxing out their credit card (and thus present the largest exposure risk). It follows that the bank could decide to lower their credit limit to mitigate such risk.

A potential future avenue of research is to more fully incorporate time to default in the prediction framework, particularly since our models confirm that max-out risk and EAD variance (dispersion) are higher the more time elapses before default. As time to default is unknown a priori, one could use survival analysis to capture its dynamic distribution, from which EAD can then be derived as in section 5.1. A follow-up study could consider different methods to implement such a PD-weighted approach (Witzany, 2011) and test their effectiveness when combined with the newly proposed EAD model. In addition, assuming the availability of a sufficiently large number of defaulted accounts, machine learning techniques, such as random forests, gradient boosting, or deep learning (depending on the structure of the data), could prove beneficial in the context of EAD modelling. Such a follow-up study could involve benchmarking a larger selection of statistical and machine learning techniques, whilst trading off any performance gains against the quality of model explanations they are able to provide.

Declaration of interest

None.

References

Barakova, I. and Parthasarathy, H. (2013). Modeling corporate exposure at default. Available at SSRN: https://ssrn.com/abstract=2235218.

BCBS (1999). *Credit risk modelling: current practices and applications*. Basel Committee on Banking Supervision. Available at https://www.bis.org/publ/bcbs49.pdf.

Brown, I. (2011). *Regression model development for credit card exposure at default (EAD) using SAS/STAT® and SAS® Enterprise Miner™ 5.3*. SAS Global Forum, Las Vegas, NV.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B -splines and penalties. *Statistical Science*, 11(2):89–121.

Gibilaro, L. and Mattarocci, G. (2018). Multiple banking relationships and exposure at default. *Journal of Financial Regulation and Compliance*, 26(1):2–19.

Gürtler, M., Hibbeln, M. T., and Usselmann, P. (2018). Exposure at default modeling – A theoretical and empirical assessment of estimation approaches and parameter choice. *Journal of Banking and Finance*, 91:176–188.

Hon, P. S. and Bellotti, T. (2016). Models and forecasts of credit card balance. *European Journal of Operational Research*, 249(2):498–505.

Jacobs, M. (2010). An empirical study of exposure at default. *Journal of Advanced Studies in Finance*, 1(1):31–59.

Kaposty, F., Löderbusch, M., and Maciag, J. (2017). Stochastic loss given default and exposure at default in a structural model of portfolio credit risk. *The Journal of Credit Risk*, 13(1):93–123.

Kupiec, P. H. (2008). A generalized single common factor model of portfolio credit risk. *The Journal of Derivatives*, 15:25–40.

Leow, M. and Crook, J. (2016). A new mixture model for the estimation of credit card exposure at default. *European Journal of Operational Research*, 249(2):487–497.

Leow, M. and Mues, C. (2012). Predicting loss given default (LGD) for residential mortgage loans: A two-stage model and empirical evidence for UK bank data. *International Journal of Forecasting*, 28(1):183–195.

Luo, S. and Murphy, A. (2020). Understanding the Exposure at Default Risk of Commercial Real Estate Construction and Land Development Loans. Working Papers 2007, Federal Reserve Bank of Dallas. Available at https://ideas.repec.org/p/fip/feddwp/87677.html.

Moral, G. (2006). EAD estimates for facilities with explicit limits. In Engelmann, B. and Rauhmeier, R., editors, *The Basel II risk parameters: Estimation, validation, and stress testing*, pages 197–242. Springer Berlin Heidelberg, Berlin, Heidelberg.

Qi, M. (2009). Exposure at default of unsecured credit cards. Economics working paper 2009-2, Office of the Comptroller of the Currency. Available at https://www.occ.gov/publications-and-resources/publications/economics/working-papers/pub-econ-working-paper-2009-2.pdf.

R Core Team (2020). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.

Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V., and De Bastiani, F. (2017). *Flexible regression and smoothing: Using GAMLSS in R.* Chapman and Hall.

Taplin, R., To, H. M., and Hee, J. (2007). Modeling exposure at default, credit conversion factors and the Basel II accord. *The Journal of Credit Risk*, 3(2):75–84.

Thackham, M. and Ma, J. (2018). Exposure at default without conversion factors – Evidence from global credit data for large corporate revolving facilities. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182(4):1267–1286.

Tong, E. N., Mues, C., Brown, I., and Thomas, L. C. (2016). Exposure at default models with and without the credit conversion factor. *European Journal of Operational Research*, 252(3):910–920.

Witzany, J. (2011). Exposure at default modeling with default intensities. *European Financial and Accounting Journal*, 6(4):20–48.

Yang, B. H. and Tkachenko, M. (2012). Modeling exposure at default and loss given default: Empirical approaches and technical implementation. *The Journal of Credit Risk*, 8(2):81–102.

Appendix: calibration plots of bucket-level EAD estimates

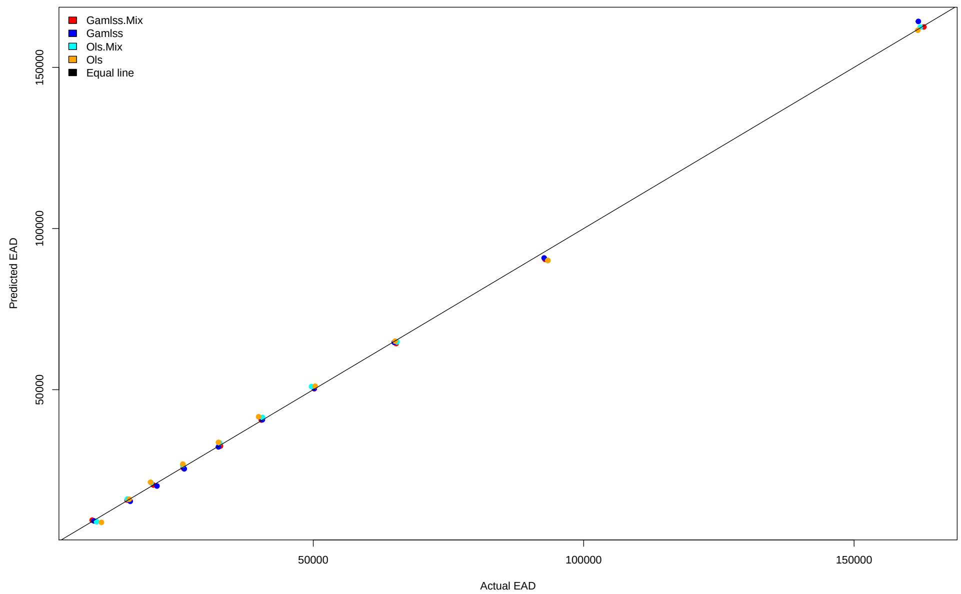

Figure 9: Calibration plots of the average of predicted EAD against that of actual EAD, assessed on different buckets of the test dataset and ordered from lowest (the leftmost point cluster) to highest (the rightmost point cluster) predicted EAD. The equal line, where the actual and predicted values are equal, is represented in diagonal. Results using actual values of time to default.

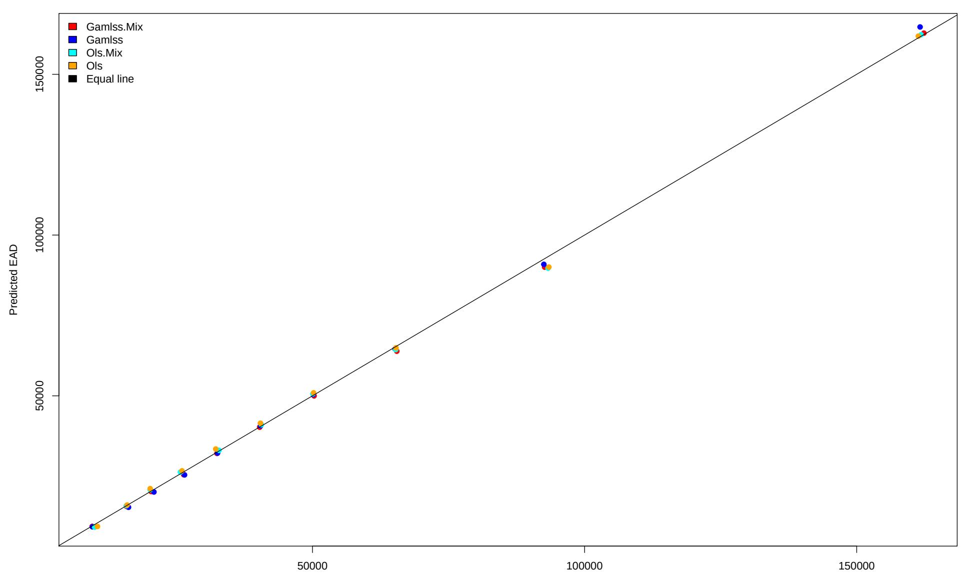

Figure 10: Calibration plots of the average of predicted EAD against that of actual EAD, assessed on different buckets of the test dataset and ordered from lowest (the leftmost point cluster) to highest (the rightmost point cluster) predicted EAD. The equal line, where the actual and predicted values are equal, is represented in diagonal. Results using the simplified PD-weighted approach.