risks

MDPI

Article

Developing an Impairment Loss Given Default Model Using Weighted Logistic Regression Illustrated on a Secured Retail Bank Portfolio

Douw Gerbrand Breed [1], Tanja Verster [1], Willem D. Schutte [1,*] and Naeem Siddiqi [2]

[1] Centre for Business Mathematics and Informatics, North-West University, Potchefstroom 2531, South Africa; gerbrand.breed@gmail.com (D.G.B.); tanja.verster@nwu.ac.za (T.V.)

[2] SAS Institute Canada, Toronto, ON M5A 1K7, Canada; naeem.siddiqi@sas.com

* Correspondence: wd.schutte@nwu.ac.za

Received: 11 October 2019; Accepted: 6 December 2019; Published: 13 December 2019

Abstract: This paper proposes a new method to model loss given default (LGD) for IFRS 9 purposes. We develop two models for the purposes of this paper—LGD1 and LGD2. The LGD1 model is applied to the non-default (performing) accounts and its empirical value based on a specified reference period using a lookup table. We also segment this across the most important variables to obtain a more granular estimate. The LGD2 model is applied to defaulted accounts and we estimate the model by means of an exposure weighted logistic regression. This newly developed LGD model is tested on a secured retail portfolio from a bank. We compare this weighted logistic regression (WLR) (under the assumption of independence) with generalised estimating equations (GEEs) to test the effects of disregarding the dependence among the repeated observations per account. When disregarding this dependence in the application of WLR, the standard errors of the parameter estimates are underestimated. However, the practical effect of this implementation in terms of model accuracy is found to be negligible. The main advantage of the newly developed methodology is the simplicity of this well-known approach, namely logistic regression of binned variables, resulting in a scorecard format.

Keywords: loss given default; weighted logistic regression; International Financial Reporting Standard 9; independence assumption

1. Introduction

The International Accounting Standard Board published the IFRS 9 standard in 2014, (IFRS 2014), which replaced most of International Accounting Standard (IAS) 39. Amongst others, it contains impairment requirements that allow for earlier recognition of credit losses. The financial statements of banks are expected to reflect the IFRS 9 accounting standards as of 1 January 2018 (European Banking Authority (EBA)). Banks found that IFRS 9 had a significant impact on systems and processes (Beerbaum 2015). While the IAS 39 standard made use of provisions on incurred losses, the financial crisis showed that expected losses, instead of incurred losses, are better used to calculate provisioning for banks (Global Public Policy Committee (GPPC)). In addition, under IFRS 9, the expected credit losses (ECL) should be equivalent to the lifetime ECL, if the credit risk has increased significantly. When the converse is true, a financial entity may allow for credit losses equal to a 12-month ECL. The ECL model is a forward-looking model and should result in the early detection of credit losses, which is anticipated to contribute to overall financial stability (IFRS 2014). The ECL is a function of the probability of default (PD), the loss given default (LGD) and the exposure at default (EAD).

In this paper, we focus on the LGD component within the impairment calculation under IFRS 9. There are many methodologies to model LGD, see e.g., Joubert et al. (2018a, 2018b) and the references therein. These methodologies include the run-off triangle method, beta regression, survival analysis, fractional response regression, inverse beta transformation, and Box–Cox transformation. Most of these techniques are quite complex and very difficult to understand, including the monitoring and validation thereof. This is confirmed by Bijak and Thomas (2018), who indicate that more than 15 different performance measures can be found in the literature concerning LGD models, possibly due to the difficulty of modelling the distribution shape of LGD. The LGD can be modelled through either the direct or the indirect approach. When using the direct approach, the LGD is equal to one minus the recovery rate (De Jongh et al. 2017). The indirect approach uses two components that are modelled separately, namely the probability component and the loss severity component. Independent of the methodology, the LGD is always assessed over the life of the lending exposure (Basel Committee on Banking Supervision 2015a).

Different modelling approaches are usually followed for accounts in different stages. An account can reside in one of three stages. Stage 1 accounts are performing accounts, Stage 2 have significant deterioration in credit risk, but are not in default, while defaulted accounts are in Stage 3 (Aptivaa 2016).

This paper describes the proposed new methodology to model the LGD for IFRS 9 purposes. We estimated both the LGD1 and LGD2 values, where the LGD1 was applied to non-defaulted accounts and the LGD2 to defaulted accounts. For the non-defaulted accounts (accounts in Stages 1 and 2, according to the IFRS 9 definition) we used the historically observed the LGD value (LGD1) and segmented this value according to variables with business importance using a lookup table. The weighted logistic regression was applied on the defaulted accounts (accounts in Stage 3, according to the IFRS 9 definition) to obtain the LGD2. This therefore resulted in two models: one for the LGD1 and one for the LGD2. The LGD1 was applied for Stage 1 (12 months) and Stage 2 (lifetime) because, while the PD component differentiates between 12 months and lifetime, the LGD is the loss expected for the remaining life of the account. Since logistic regression is well known and regularly used in banks, established monitoring metrics and governance practices have been embedded in the industry. These metrics, as well as the methodology, are thoroughly understood by stakeholders, which leads to a high degree of confidence in the results. Logistic regression using the scorecard format provides an even more transparent and user-friendly technique that is easy to understand and communicate to stakeholders. For this reason, we propose this new methodology to model the LGD for IFRS 9 purposes.

The paper consists of five sections. The modelling approach is described in Section 2. Section 3 follows with a case study where the proposed methodology is applied to a secured retail portfolio. The effect of the dependency of the observations used in the logistic regression is tested by comparing the results from the logistic regression with that of a generalised estimating equation (that takes dependency into account). We also investigate whether a decision tree could outperform the weighted logistic regression. Section 4 discusses the strengths and weaknesses of our new methodology and Section 5 concludes.

2. LGD Methodology

This section describes the newly proposed LGD methodology. First, the methodology used to estimate the LGD of the non-defaulted accounts is provided (LGD1) under Section 2.1, followed by Section 2.2 that discusses the methodology employed to model the LGD of the defaulted accounts (LGD2).

2.1. LGD1 Methodology

The LGD1 was obtained by calculating the loss as the exposure at default minus the net present value (NPV) of recoveries, divided by the EAD. The LGD1 is typically modelled on a smaller data sample than for the LGD2, since the loss is only calculated on accounts in Stages 1 and 2 that eventually transition into default. The probability of transitioning from Stages 1 and 2 directly into default is typically very low. The data sample for LGD2 is typically much larger as it considers all accounts in default and not only those that transition into it in a specific cohort. In this specific case study, the discounted write-off amount served as a proxy for the NPV of recovery cash flows. A more granular estimate was obtained by using the most important LGD drivers to segment the LGD1 values. The historical LGD values were calculated (and averaged) per segment and the estimated LGD1 values were derived using a lookup table (see e.g., Basel Committee on Banking Supervision (2015b)). The results of the case study are shown in Section 3.1. Note that the number of variables available to use for the LGD2 modelling is much larger than that for the LGD1. The reason is that some of the default related variables are not available at the LGD1 model stage, e.g., months since default.

2.2. LGD2 Methodology

We use a weighted logistic regression to model the LGD for the defaulted account including all available data. The actual loss experienced is transformed to a binary format related to the "fuzzy augmentation" technique commonly used to introduce "rejects" in scorecard development (Siddiqi 2006). This means that each observation has both a target of 1 (Y = 1) as well as a target of 0 (Y = 0). Furthermore, a weight variable is created, where the sum of the weights of these two events adds up to the full exposure of the account at observation. This is related to Van Berkel and Siddiqi (2012) who used a scorecard format for modelling LGD. This newly proposed methodology for LGD2 only considers worked-out accounts. A worked-out account can either cure or be written off. Note that the point of write-off is taken as that specific point where the institution (e.g., bank) no longer expects any recovery. This is specifically prescribed by the reporting standard: "IFRS 7 (35F) (e): The Group writes off financial assets, in whole or in part, when it has exhausted all practical recovery efforts and has concluded there is no reasonable expectation of recovery. Indicators that there is no reasonable expectation of recovery include (i) ceasing enforcement activity and (ii) where the Group's effort to dispose of repossessed collateral is such that there is no reasonable expectation of recovering in full" (PWC 2017). In effect, with our methodology, all write-offs and cures are included regardless of the time spent in default and no filter is applied on default cohort.

We calculated the LGD for accounts that cured and for accounts that were written off. The modelling approach can be subdivided into five steps: (1) sample creation; (2) target and weight variables created; (3) input variables; (4) weighted logistic regression; and (5) test for independence.

Note that if any accounts in the considered dataset originated as credit impaired accounts (i.e., accounts starting in default), their loss behaviour will most likely be different from other Stage 3 accounts and should therefore be modelled separately (i.e., segment the portfolio based on this characteristic). In this specific case study here presented, no such accounts existed.

2.2.1. Step 1: Sample Created

This approach would first need to identify all worked-out accounts (i.e., write-off, cure) over an appropriate reference period. Note that one account can appear multiple times, but an account will only appear once per month. This violates the logistic regression assumption of independent observations. The effect of this dependence (Sheu 2000) is tested at the end of the paper, by comparing a generalised estimating equation (GEE) model with the logistic regression model. In statistics, a GEE is used to estimate the parameters of a generalised linear model with a possible unknown correlation between outcomes (Kuchibhatla and Fillenbaum 2003).

The sample of observations was split into two datasets for out of sample testing. Evaluating the performance of a classifier on the same data used to train the classifier usually leads to an optimistically biased assessment (SAS Institute 2010). The simplest strategy for correcting the optimism bias is to hold out a portion of the development data for assessment (Baesens et al. 2016), i.e., data splitting. We therefore split the data into a training and a validation dataset. The validation data is used only for assessment and not for model development.

2.2.2. Step 2: Target and Weight Variables Created

Two rows ($Y = 1$ and $Y = 0$) are created for each observation (i.e., per account per month). Each row is weighted. Cured and written-off accounts are weighted differently. Mathematically, the weight for observation i is defined as

$$w_i = \begin{cases} Exposure_i \times LGD_i & if\ Y_i = 1 \\ Exposure_i \times (1 - LGD_i) & if\ Y_i = 0, \end{cases} \tag{1}$$

where the loss given default of observation i (LGD_i) is defined as

$$LGD_i = \begin{cases} P(Cure) \times P(redefault) \times LGD_{1.Unadj} & \text{if observation } i \text{ is a cured} \\ WO_i / Exposure_i & \text{if observation } i \text{ is written off} \end{cases}, \tag{2}$$

where

- i is the number of observations from 1 to N;
- $Exposure_i$ is the exposure of observation i; and therefore,

$$EAD_i = \sum_{\forall Y_i} Exposure_i = Exposure_i \text{IND}(Y_i = 1) + Exposure_i \text{IND}(Y_i = 0), \tag{3}$$

where $IND(Y_i = 1) := \begin{cases} 1\ if\ Y_i = 1 \\ 0\ if\ Y_i = 0 \end{cases}$ and $IND(Y_i = 0) := \begin{cases} 1\ if\ Y_i = 0 \\ 0\ if\ Y_i = 1 \end{cases}$.

- $P(Cure)$ is the proportion of cured observations over the total number of worked-out accounts (over the reference period);
- $P(redefault)$ is the proportion of observations that re-default over the reference period;
- $LGD_{1.Unadj}$ is the exposure at default (EAD) minus the net present value (NPV) of recoveries from first point of default for all observations in the reference period divided by the EAD—see e.g., PWC (2017) and Volarević and Varović (2018);
- WO_i is the discounted write-off amount for observation i; and
- $P(Cure)$, $P(redefault)$ and $LGD_{1.Unadj}$ are therefore empirical calculated values. This should be regularly updated to ensure the final LGD estimate remains a point in time estimate as required by IFRS (IFRS 2014).

Note that the write-off amount is used in Equation (2) to calculate the actual LGD. An alternative method employs the recovery cash flows over the work out period. A bank is required to use its "best estimate" (a regulatory term, e.g., Basel Committee on Banking Supervision (2019b) and European Central Bank (2018)) to determine actual the LGD. In this case, this decision was based on the data available. Only the write-off amount was available for our case study, not the recovered cash flows. In Equation (2), the write-off amount needs to be discounted using the effective interest rate (PWC 2017), to incorporate time value of money. When recoveries are used, each recovery cash flow needs to be similarly discounted. In the case study, the length of the recovery time period exists in the data and differs for each account. The length of this recovery time period will have an influence on the calculation of LGD: the longer the recovery process, the higher the effective discount rate. In the case study, we used the client interest rate as the effective interest rate when discounting.

Note that, in special circumstances, accounts may be partially written off, leading to an overestimation of provision. This should be taken into account during the modelling process. However, in our case study no such accounts existed.

Illustrative Example

Consider one observation with an exposure of $50,000. Assume it is a written-off account, for a specific month, with an LGD_i = 27% (based on the written-off amount divided by the exposure, i.e.,

$WO_i/Exposure_i$). The weight variable for $Y = 1$ will be 27% × \$50,000 = \$13,500 and $Y = 0$ will be (1 − 27%) × \$50,000 = \$36,500 (see Table 1).

Table 1. Illustrative example of weight variable.

Binary Outcome (Y)	Exposure	Weight Variable
0	\$50,000	\$13,500
1	\$50,000	\$36,500

2.2.3. Step 3: Input Variables (i.e., Variable Selection)

All input variables were first screened according to the following three requirements: percentage of missing values, the Gini statistic and business input. If too many values of a specific variable were missing that variable was excluded. Similarly, if a variable had a too low value for the Gini statistic, then that variable was also excluded. Note that business analysts should investigate whether there are any data issues with variables that have low Gini statistics. For example, traditionally strong variables may appear weak if the data has significant sample bias. This forms part of data preparation that is always essential before predictive modelling should take place.

The Gini statistic (Siddiqi 2006) quantifies a model's ability to discriminate between two possible values of a binary target variable (Tevet 2013). Cases are ranked according to the predictions and the Gini then provides a measure of correctness. It is one of the most popular measures used in retail credit scoring (Baesens et al. 2016; Siddiqi 2006; Anderson 2007) and has the added advantage that it is a single value (Tevet 2013).

The Gini is calculated as follows (SAS Institute 2017):

1. Sort the data by descending order of the proportion of events in each attribute. Suppose a characteristic has m attributes. Then, the sorted attributes are placed in groups $1, 2, \ldots, m$. Each group corresponds to an attribute.
2. For each of these sorted groups, compute the number of events $\left((\#(Y = 1)_j\right)$ and the number of nonevents (#(Y=0)_j)in group j. Then compute the Gini statistic:

$$\left(1 - \frac{2\sum_{j=2}^{m}\left((\#(Y=1)_j \times \sum_{j=1}^{j-1}\#(Y=0)_j\right) + \sum_{j=1}^{m}\left((\#(Y=1)_j \times \#(Y=0)_j\right)}{\#(Y=1) \times \#(Y=0)} \times 100\right), \tag{4}$$

where $\#(Y = 1)$ and $\#(Y = 0)$ are the total number of events and nonevents in the data, respectively.

Only variables of sufficient Gini and which were considered important from a business perspective were included in the modelling process. All the remaining variables after the initial screening were then binned. The concept of binning is known by different names such as discretisation, classing, categorisation, grouping and quantification (Verster 2018). For simplicity we use the term binning throughout this paper. Binning is the mapping of continuous or categorical data into discrete bins (Nguyen et al. 2014). It is a frequently used pre-processing step in predictive modelling and considered a basic data preparation step in building a credit scorecard (Thomas 2009). Credit scorecards are convenient points-based models that predict binary events and are broadly used due to their simplicity and ease of use; see e.g., Thomas (2009) and Siddiqi (2006). Among the practical advantages of binning are the removal of the effects of outliers and a convenient way to handle missing values (Anderson 2007). The binning was iteratively done by first generating equal-width bins, followed by business input-based adjustments to obtain the final set. Note that if binned variables are used in logistic regression, the final model can easily be transformed into a scorecard.

All bins were quantified by means of the average LGD value per bin. The motivation behind this was to propose an alternative to using dummy variables. Logistic regression cannot use categorical variables coded in its original format (Neter et al. 1996). As such, some other measure is needed for each bin to make it usable—the default technique of logistic regression is a dummy variable for each class less one. However, expanding categorical inputs into dummy variables can greatly increase the dimension of the input space (SAS Institute 2010). One alternative to this is to quantify (e.g., using weights of evidence (WOE)—see Siddiqi (2006)) each bin using the target value (in our case the LGD value), which will reduce the number of estimates. An example of this is using the natural logarithm (ln) of the good/bad odds (i.e., the WOE)—see for example Lund and Raimi (2012). We used the standardised average LGD value in each bin.

Some of the advantages of binning and quantifying the bins are as follows:

- The average LGD value can be calculated for missing values, which will allow ″Missing″ to be used in model fit (otherwise these rows would not have been used in modelling). Note that not all missing values are equal and there are cases where they need to be treated separately based on reason for missing, e.g., "No hit" at the bureau vs. no trades present. It is therefore essential that business analysts investigate the reason for missing values and treat them appropriately. This again forms part of data preparation that is always a key prerequisite to predictive modelling.
- Sparse outliers will not have an effect on the fit of the model. These outliers will become incorporated into the nearest bin and their contributions diminished through the usage of bin WOE or average LGD.
- Binning can capture some of the generalisation (required in predictive modelling). Generalisation refers to the ability to predict the target of new cases and binning improves the balance between being too vague or too specific.
- The binning can capture possible non-linear trends (as long as they can be assigned logical causality).
- Using the standardised average LGD value for each bin ensures that all variables are of the same scale (i.e., average LGD value).
- Using the average LGD value ensures that all types of variables (categorical, numerical, nominal, ordinal) will be transformed into the same measurement type.
- Quantifying the bins (rather than using dummy variables) results in each variable being seen as one group (and not each level as a different variable). This aids in reducing the number of parameter estimates.

Next, each of these average LGD values was standardised using the weight variable by calculating the average LGD per bin. An alternative approach could have been to calculate the WOE for each bin. The WOE is regularly used in credit scorecard development (Siddiqi 2006) and is calculated using only the number of 1′s and the number of 0′s for each bin. Note that our underlying variable of interest (LGD) is continuous. However, since our modelled target variable was dichotomous, we wanted the quantification of the bin to reflect our underlying true target, e.g., the LGD value, which ranges from 0 to 1. This average LGD value per bin was then standardised by means of the weight variable. The weighted mean LGD, $\overline{LGD}_w$ is defined as

$$\overline{LGD}_w = \frac{\sum_i w_i LGD_i}{\sum_i w_i}, \tag{5}$$

where LGD_i is the LGD value of observation i and w_i is the weight of observation i. The weighted standard deviation LGD is defined as

$$s_w = \sqrt{\frac{\sum_i w_i \left(LGD_i - \overline{LGD}_w\right)^2}{N-1}}, \tag{6}$$

where N is the number of observations. The weighted standardised value for LGD, $LGD^{*}{}_{i}$, for observation i will then be

$$LGD^{*}{}_{i} = \frac{LGD_i - \overline{LGD}_w}{s_w}. \tag{7}$$

The standardisation of all input variables implies that the estimates from the logistic regression will be standardised estimates. The benefit is that the absolute value of the standardised estimates can serve to provide an approximate ranking of the relative importance of the input variables on the fitted logistic model (SAS Institute 2010). If this was not done, the scale of each variable could also have had an influence on the estimate. Note that the logistic regression fitted was a weighted logistic regression with the exposure as weight (split for $Y = 1$ and $Y = 0$) and therefore to ensure consistency, we also weighted the LGD with the same weight variable as used in the logistic regression.

Furthermore, pertaining to the month since default as input variable: The model that is developed does not require the length of default for incomplete accounts in order to estimate LGD. It assumes that the length of default for these accounts will be comparable to similar accounts that have been resolved. This is an assumption that can be easily monitored after implementation.

2.2.4. Step 4: Weighted Logistic Regression

A weighted logistic regression was then fitted using the available data. The log of the odds in a weighted logistic regression is given as:

$$logit(p_i) = \ln\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \boldsymbol{\beta} w_i \boldsymbol{X}_i^T, \tag{8}$$

where:

- $p_i = E(Y_i = 1 \middle| \boldsymbol{X}_i, \beta)$ is the probability of loss for observation i;
- β_0, $\boldsymbol{\beta}$ are regression coefficients with $\boldsymbol{\beta} = \{\beta_{1,}\ \ldots,\ \beta_K\}$;
- $\boldsymbol{X}_i$ is the vector of the predictor variables $X_{i1,}, \ldots, X_{iK}$ for observation i; and
- w_i is the weight of each observation i, calculated by the actual loss amount (\$s) and given in Equation (1).

Note that in this weighted logistic regression, we estimated the regression coefficients where the repeated observation within the same individual was assumed to be independent (i.e., disregarding the dependence among repeated observation of the same account).

2.2.5. Step 5: Test the Effect of the Dependence Assumption

A single account would appear multiple times in our dataset (depending on the number of months the account is present), which violates the assumption of independent observations in logistic regression. We therefore tested the effect of this violation using a GEE that can handle the statistical dependence of repeated data by assuming some correlation structure (Kuchibhatla and Fillenbaum 2003) among observations. This approach estimates regression coefficients without completely specifying the joint distribution of the multivariate responses, but the parameters of the within-subjects correlation are explicitly accounted for in the estimation process (Sheu 2000). It is also shown in Sheu (2000) that the GEE approach is not sensitive to the choice of correlated structure. Kuchibhatla and Fillenbaum (2003) also found that when comparing the model fit using the GEE with that using the logistic regression, the logistic regression overestimated the standard errors of the dependent variables.

3. Case Study: Secured Retail Portfolio from a South African Bank

This section illustrates the newly proposed LGD methodology on a secured retail portfolio from one of the major banks in South Africa. Section 3.1 shows the results for the non-defaulted accounts (i.e., LGD1). Then, Section 3.2 shows results for the defaulted accounts (LGD2). Note that the data was split into LGD1 and LGD2, resulting in 95% of the data in the LGD2 dataset and 5% of the data in the LGD1 dataset. The reason for the much smaller LGD1 dataset is that very few of the "non-defaulted" sub-set of total accounts actually defaulted. In reality, the LGD1 model is applied to the non-defaulted portfolio (which is typically the bigger dataset), whereas the LGD2 model is applied to the defaulted portfolio (which is typically a much smaller dataset). The datasets used for modelling therefore appear counterintuitive to real world conditions.

3.1. LGD1 Results

The empirical observed LGD described in Section 2.1 was applied to the pre-default book, i.e., accounts not in default. This number was further enhanced by using segmentation variables. As it is important that the variables used for segmentation do not change over time, the final set of variables was selected on the basis of stability and business sense. These variables were then binned. The final variables selected for segmentation were loan to value (LTV) at origination, channel/manufacturer and new/old/used indicator (NOU). The channel/manufacturer variable was derived the channel and manufacturer code. The empirical LGDs at the point of default were subsequently calculated by these variables in a matrix type approach (lookup table). The final lookup table for accounts not in default (Stage 1 and 2) is in Table 2. Note that the standardised LGD values are shown to protect the confidential information surrounding this portfolio's observed values. The final segmentation is consistent with business sense (note the LGD separation from very negative to very positive). This lookup table approach—separating risks into different bins (slots)—is closely related to the concept of slotting (Basel Committee on Banking Supervision 2019a).

Table 2. Lookup table for Loss Given Default 1.

LTV	Channel & Manufacturer	New/Old	Standardised LGD
<=1	Group 1	New	−1.0553
<=1	Group 1	Old	−1.00075
<=1	Group 2	New	−0.87389
<=1	Group 2	Old	−0.18252
<=1	Group 1	New	−0.2155
<=1	Group 1	Old	−0.10513
<=1	Group 3	New	−0.67346
<=1	Group 3	Old	0.050902
>1	Group 1	New	−0.22311
>1	Group 1	Old	0.519007
>1	Group 2	New	−0.24721
>1	Group 2	Old	0.532962
>1	Group 1	New	0.365509
>1	Group 1	Old	0.957936
>1	Group 3	New	0.647134
>1	Group 3	Old	1.503425

3.2. LGD2 Results

The results are described according to the five steps discussed in Section 2.2.

3.2.1. Step 1: Sample Created

A 24-month reference period, based on business input, was used. Only worked-out accounts were selected in this reference period. The LGD2 dataset was then split into a 70% training (946,285 observations, 38,352 unique accounts) and 30% validation dataset (405,630 observations, 37,720 unique accounts).

3.2.2. Step 2: Target and Weight Variables Created

Two rows ($Y = 1$ and $Y = 0$) were created for each observation (i.e., per account per month). Each row was weighted, as described in Section 2.2.

3.2.3. Step 3: Input Variables

All input variables were screened using the following three requirements: percentage of missing values (more than 50% missing was used as a cut-off), the Gini statistic (variables with low Gini statistic values were excluded) and business input. All bins were quantified using the average LGD value per bin, which was then standardised with the weight variable. Table 3 lists the binned and quantified variables used in the weighted logistic regression. The final decision on binning was a combination of bucket stability (CSI), logical trends as well as consistency of logical trends over time. Some of the observations on these variables, with respect to LGD values, include (variable names indicated in brackets):

- Higher LTV values are associated with higher LGD values (*LTV*).
- The higher the month on book (MOB) value for a customer, the lower the expected LGD value (*MOB*).
- The more months a customer has been in default, the higher the LGD value (*Default*).
- Customers buying old vehicles are associated with higher LGD values (*New/Old*).
- Certain channels and certain manufacturers are associated with higher LGD values (*Channel Manufacturer*).

Table 3. Binned input variables.

LTV	**LTV Range**	**#**	**Standardised LGD**
Bin 1	LTV <=1	18188	−0.00566
Bin 2	LTV <=1.2	10461	−0.00268
Bin 3	LTV > 1.2	9703	0.004802
MOB	**MOB Range**	**#**	**Standardised LGD**
Bin 1	MOB <=24	17593	0.005193244
Bin 2	MOB <=42	10431	−0.000342394
Bin 3	MOB > 42	10328	−0.006198457
Default	**Default Range**	**#**	**Standardised LGD**
Bin 1	0	1043	−0.005747327
Bin 2	1	7706	−0.004411893
Bin 3	2+	16150	−0.000289465
Bin 4	Other/Missing	13453	0.006032881
New/Old	**New/Old Range**	**#**	**Standardised LGD**
Bin 1	New	15249	−0.004677389
Bin 2	Old	23103	0.004428005
Channel Manufacturer	**Channel Manufacturer Range**	**#**	**Standardised LGD**
Bin 1	Group 1	3870	−0.008325
Bin 2	Group 2	5984	−0.004694
Bin 3	Group 3	26422	0.001172
Bin 4	Group 4	2076	0.011212

This binning approach (separating risks in different bins or slots) is related to the underlying principle used in slotting (Basel Committee on Banking Supervision 2019a).

3.2.4. Step 4: Weighted Logistic Regression

A stepwise weighted logistic regression was fitted on the dataset, with a 5% significance level. The analyses were performed using SAS. The SAS code is provided as Supplementary Material for reference. While the authors of this paper used SAS, users can implement these techniques using any available analytic tool including Python and R. The final variables for accounts in default (Stage 3) are given in Table 4. The Gini statistic on the training dataset was 45.49% and on the validation dataset 36.04%. The difference between the training and validation Gini is quite large and could be an indication of the model not generalising well. However, it should be acknowledged that the Gini describes how well the model distinguishes between the two groups $Y_i = 1$ and $Y_i = 0$ (Breed and Verster 2017), while our underlying target is the LGD value which is a continuous value between 0 and 1. Therefore, a better measure for both model performance and comparing training and validation samples is the mean squared error (MSE), although several other measures could have been used (see Bijak and Thomas (2018) for an extensive list of performance measures applicable to LGD).

Table 4. Weighted logistic regression results.

Analysis of Maximum Likelihood Estimates					
Parameter (X)	**DF**	**Estimate (β)**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	−1.0977	0.000012	8528907254	<0.0001
LTV (X_1)	1	32.6329	0.00256	161977546	<0.0001
Months on books (X_2)	1	10.3046	0.00261	15622966.5	<0.0001
Default event (X_3)	1	173.9	0.00253	4709270394	<0.0001
New/Old (X_4)	1	18.5934	0.00252	54593987.2	<0.0001
Channel/Manufacturer (X_5)	1	17.3602	0.00248	48935118.5	<0.0001

The mean squared error was therefore calculated as follows:

$$MSE_i = \frac{\left(\widehat{LGD}_i - LGD_i\right)^2}{N}, \tag{9}$$

where $\widehat{LGD}_i$ is the predicted LGD value of observation i from the model, LGD_i the best estimate of the actual LGD (as defined in Equation (1)) and where i is the number of observations from 1 to N.

The MSE for the training and validation datasets were 0.0473 and 0.0427 respectively, thus showing a small mean squared error.

The R-square value (Bijak and Thomas 2018) was calculated as follows:

$$R\ squared = 1 - \frac{\sum_i\left(\widehat{LGD}_i - LGD_i\right)^2}{\sum_i\left(\widehat{LGD}_i - \overline{\mathrm{LGD}}\right)^2}, \tag{10}$$

where $\overline{LGD}$ is the expected value of the actual LGD values. The R-squared value for the training and validation datasets were 0.3202 and 0.2727, respectively.

Furthermore, it showed that the model generalises well, as it also predicted well on the validation dataset (small difference between train and validation datasets). Here, the MSE on the training and validation dataset are very close. Note that the MSE and the R-squared value were calculated on the LGD values and not the standardised LGD values.

3.2.5. Step 5: Test the Effect of the Dependence Assumption

Next, we estimated the regression coefficients using the GEE, first assuming an independent correlation structure and then with an autoregressive correlation structure of the order one. In Table 5, the results of the GEE using an independent correlation structure are shown. The code is provided in the Supplementary Material for reference, although any other analytics tool could be used to fit a GEE model.

Table 5. Generalised Estimating Equations regression results (independent correlation).

Analysis of GEE Parameter Estimates						
Empirical Standard Error Estimates						
Parameter (X)	**Estimate (β)**	**Standard Error**	**95% Confidence Limits**		**Z**	**Pr > |Z|**
Intercept	−1.0978	0.0116	−1.1205	−1.0750	−94.44	<0.0001
LTV (X_1)	32.6348	2.4257	27.8805	37.3891	13.45	<0.0001
Months on books (X_2)	10.3055	2.3708	5.6587	14.9522	4.35	<0.0001
Default event (X_3)	173.8758	1.8297	170.2897	177.4619	95.03	<0.0001
New/Old (X_4)	18.5943	2.4984	13.6976	23.4910	7.44	<0.0001
Channel Manufacturer (X_5)	17.3607	2.5861	12.2921	22.4293	6.71	<0.0001

The Gini on the training and validation datasets came to 45.49% and 36.04%, respectively, while the MSE on the training and validation datasets were 0.0473 and 0.0427. We used a significance level of 5% throughout, and all six variables were statistically significant. We note that the results (parameter estimates, Gini and MSE) are almost identical to that of the weighted logistic regression.

Next, an autoregressive correlation structure to the order of one was assumed. The results are shown in Table 6.

Table 6. Generalised Estimating Equations regression results (autoregressive correlation).

Analysis of GEE Parameter Estimates						
Empirical Standard Error Estimates						
Parameter (X)	**Estimate (β)**	**Standard Error**	**95% Confidence Limits**		**Z**	**Pr > |Z|**
Intercept	−0.7973	0.0080	−0.8131	−0.7816	−99.15	<0.0001
LTV (X_1)	24.8404	1.8335	21.2468	28.4339	13.55	<0.0001
Months on books (X_2)	6.8528	1.7314	3.4592	10.2463	3.96	<0.0001
Default event (X_3)	129.6377	1.3393	127.0126	132.2627	96.79	<0.0001
New/Old (X_4)	12.5228	1.8139	8.9677	16.0779	6.90	<0.0001
Channel Manufacturer (X_5)	11.7312	1.8959	8.0154	15.4470	6.19	<0.0001

The Gini on the training data was 45.48% and on the validation dataset 36.04%, with the MSE values being 0.0522 and 0.0406, respectively, for training and validation. Note that all six variables were again statistically significant.

Next, the three models were compared in terms of parameter estimates, standard errors of the estimates and on model performance. Table 7 provides the parameter estimates comparisons, which indicate similar numbers for the weighted logistic regression (LR) and the GEE (independent correlation). This is similar to the results found by Sheu (2000). The parameter estimates were quite different, however, when using an autoregressive correlation matrix.

Table 7. Comparison of the parameter estimates of the three modelling techniques.

	Weighted LR	**GEE (Ind Corr)**	**GEE (Ar1 Corr)**
β_0	−1.0977	−1.0978	−0.7973
β_1	32.6329	32.6348	24.8404
β_2	10.3046	10.3055	6.8528
β_3	173.9	173.8758	129.6377
β_4	18.5934	18.5943	12.5228
β_5	17.3602	17.3607	11.7312

In Table 8 the most significant difference between using a weighted logistic regression (disregarding the dependence among repeated observations of the same account) and using a GEE (addressing the dependence) can be seen. The weighted logistic regression underestimates the standard error of the parameter estimates. This is also confirmed by Sheu (2000) and Kuchibhatla and Fillenbaum (2003). Disregarding the dependence leads to the incorrect estimation of the standard errors. Although this is a problem from a statistical standpoint, resulting in incorrect inferences of the parameters, the practical effect is negligible, as evident from the goodness-of-fit statistics (MSE) of the different models.

Table 8. Comparison of the standard errors of the three modelling techniques

	Weighted LR	GEE (Ind Corr)	GEE (Ar1 Corr)
β_0	0.000012	0.0116	0.0080
β_1	0.00256	2.4257	1.8335
β_2	0.00261	2.3708	1.7314
β_3	0.00253	1.8297	1.3393
β_4	0.00252	2.4984	1.8139
β_5	0.00248	2.5861	1.8959

Table 9 summarises the model performance of all three models. It is interesting to note that all three models have almost identical performance. From a practical point of view, there was no difference in using any of these three techniques. When the model is productionalised, the bank will use the model to predict a specific LGD value and the accuracy of this predicted LGD was almost identical with either technique. If we suppose that the standard errors are not used by the bank, then there is no reason to refrain from using logistic regression.

Table 9. Comparison of the model performance of the three modelling techniques.

Technique	Train MSE	Valid MSE	Train Gini	Valid Gini
Weighted logistic regression	0.04727719	0.04274367	0.45492145910	0.36039085030
GEE (independent correlation)	0.04727703	0.04274417	0.45492145910	0.36039085030
GEE (AR 1 correlation)	0.05222953	0.04062386	0.45482289180	0.36037450660

One additional issue that bears mentioning is the low number of variables in the model itself. The banking industry prefers to see models that are consistent with how they would make decisions—meaning models must have variables that not only make business sense, but also cover as many of the different information types that should be considered. Typically, between eight to fifteen variables are considered normal in the industry (Siddiqi 2017). In a business setting, it is also common to add weaker variables, albeit those that display satisfactory correlations with the target, into the model itself.

3.3. Additional Investigation: Decision Tree

An additional modelling technique, namely the decision tree (Breiman et al. 1984), was considered to determine whether it could improve on the results above. First, the distribution of the actual LGD was analysed, as shown in Figure 1 (training dataset). Note that the LGD values were standardised, by subtracting the average and then dividing by the standard deviation. It can be seen that the LGD has a huge spike to the left and a much smaller spike closer to the right. This bimodal type of distribution is typical of an LGD distribution (Joubert et al. 2018a).

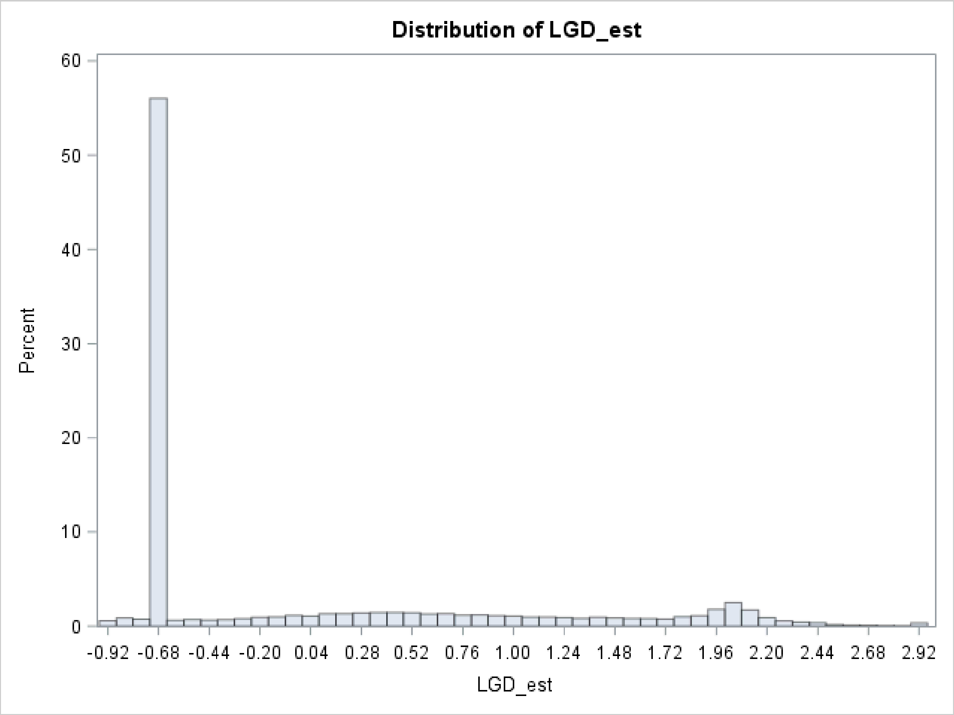

Figure 1. The distribution of the standardised observed LGD (training dataset).

The decision tree (i.e., classification tree), was developed with three different settings (see the Supplementary Material for the specific code used). First, the default settings were used. Second, the decision tree was pruned by means of the average squared error (ASE), and lastly, a setting of "no pruning" was used. For each of the decision trees, we used the same target variable (binary variable), the same weight variable and the same six explanatory variables as with the other models developed in this section. The MSE (Table 10) for the decision tree was worse than the weighted logistic regression and the GEE models that were developed (Table 9). Note that we only show the MSE values here and not the Gini values, because, as noted before, the MSE is a better measure to indicate model performance of the "true" target value, i.e., the LGD values.

Table 10. Model performance of decision trees.

Technique	Valid MSE
Decision tree (default settings)	0.1012884759
Decision tree (prune on ASE)	0.1002412789
Decision tree (no pruning)	0.1041756997

4. Strengths and Weaknesses of the Methodology

The method discussed in this paper presents several advantages. The first is that it is a relatively simplistic approach. Logistic regression is a well-known technique that has a long history in the financial services industry. In contrast, for secured products, indirect more complex methodologies are often used. One example is using a haircut model for the loss severity component and survival analysis for the probability component (Joubert et al. 2018b). Because logistic regression is well known and regularly used in banks, established monitoring metrics and governance practices have been embedded in the industry. These metrics, as well as the methodology, are thoroughly understood by stakeholders, which leads to a high degree of confidence in the results. Logistic regression using the scorecard format provides an even more transparent and user-friendly technique that is easy to understand and communicate to stakeholders.

A second advantage is that all variables are first binned and then quantified using the standardised LGD rate in each bin. Some of the specific advantages of this type of data transformation, as noted earlier in the paper, are:

- Better handling of missing values, and their usage in the model.
- Better way to deal with outliers by minimising their influence.
- Improved generalisation of data.
- Easier way to capture non-linear trends.
- Easier comparison across variables through the usage of standardised average LGD value for each bin and standardised estimates.
- A reduction in the degrees of freedom introduces stability into the model.

A weakness of the weighted regression is that it disregards the assumption of independence and this results in the statistical inference of the parameter estimates being incorrect. In particular, the standard errors of the parameter estimates are underestimated. Yet, there is no apparent difference in model accuracy.

5. Conclusions and Recommendation

This paper presented a new methodology to model LGD for IFRS 9 purposes, consisting of two components. First, the LGD1 model was applied to the non-default accounts and is an empirical value obtained through a lookup table, based on a specified reference period. This LGD1 was further segmented across the most important variables to obtain a more granular estimate. Second, the LGD2 was applied to defaulted accounts and is estimated using an exposure weighted logistic regression. This new methodology was tested by applying it on a real dataset, using a secured retail bank portfolio.

A comparison of this weighted logistic regression was done with GEE models to test the effect of the dependence among repeated observation of the same account. We discovered that when disregarding the repeated accounts, the standard errors of the parameter estimates were underestimated. However, the practical effects of such disregard were found to be negligible.

In conclusion, we propose this new methodology to model LGD for IFRS 9 purposes based on the following reasons mentioned in the paper:

- This methodology presents a relatively simple approach using logistic regression, which is a well-known and accepted technique in the banking industry.
- The results are easy to interpret and understand, and when converted to the scorecard format, provide a transparent user-friendly output.
- The method also uses transformations that offer better alternatives for dealing with issues such as missing data and outliers.
- Most banks have well-established processes for monitoring and implementing logistic regression models and they are well understood by stakeholders.
- From a practical perspective, there was no discernible difference in model accuracy when comparing the logistic regression model to the GEE model or the decision tree.

From a purely theoretical point of view, we recommend using the GEE approach. However, as some banks do not use the parameter estimates or the associated standard errors for any decisions (e.g., variable selection), the weighted logistic regression approach may be preferable in such situations.

We suggest future research ideas to include comparing this new methodology to other LGD modelling techniques. We could also explore alternative data transformations from the current binning and quantification using standardised LGD rates. We also did not include any direct costs in the calculation of the LGD, and determining how to split costs into direct and indirect components could be a further research idea. According to IFRS 9, the LGD should include forward-looking macro-economic scenarios (Miu and Ozdemir 2017). This has also not been considered in this paper and could be researched in future.

Supplementary Materials: The following are available online at http://www.mdpi.com/2227-9091/7/4/123/s1, The complete SAS code used in this publication.

Author Contributions: Conceptualization, D.G.B., T.V. and W.D.S.; formal analysis, D.G.B., T.V. and W.D.S.; investigation, D.G.B., T.V. and W.D.S.; methodology, D.G.B., T.V., W.D.S. and N.S.; software, D.G.B., T.V. and W.D.S.; validation, D.G.B., T.V., W.D.S. and N.S.; visualization, D.G.B., T.V. and W.D.S.; writing—original draft, D.G.B., T.V. and W.D.S.; writing—review and editing, D.G.B., T.V., W.D.S. and N.S.

Funding: This research received no external funding.

Acknowledgments: This work is based on research supported in part by the Department of Science and Technology (DST) of South Africa. The grant holders at the Centre for Business Mathematics and Informatics acknowledges that opinions, findings and conclusions or recommendations expressed in any publication generated by DST-supported research are those of the author(s) and that the DST accepts no liability whatsoever in this regard.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

Anderson, Raymond. 2007. *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*. Oxford: Oxford University Press.

Aptivaa. 2016. Cash Shortfall & LGD – Two Sides of the Same Coin. Available online: http://www.aptivaa.com/blog/cash-shortfall-lgd-two-sides-of-the-same-coin/ (accessed on 4 March 2019).

Baesens, Bar, Daniel Rosch, and Harald Scheule. 2016. *Credit Risk Analytics*. Cary: SAS Institute, Wiley.

Basel Committee on Banking Supervision. 2015a. Guidance on Accounting for Expected Credit Losses. Bank for International Settlements. Available online: https://www.bis.org/bcbs/publ/d350.htm (accessed on 31 January 2017).

Basel Committee on Banking Supervision. 2015b. Revisions to the Standardised Approach for Credit Risk. Bank for International Settlements. Available online: https://www.bis.org/bcbs/publ/d347.pdf (accessed on 18 February 2019).

Basel Committee on Banking Supervision. 2019a. CRE33 IRB Approach: Supervisory Slotting Approach for Specialised Lending (CRE Calculation of RWA for Credit Risk). Bank for International Settlements. Available online: https://www.bis.org/basel_framework/chapter/CRE/33.htm?tldate=20220101&inforce=20190101&export=pdf&pdfid=15661993943265707 (accessed on 11 March 2019).

Basel Committe on Banking Supervision. 2019b. Calculation of RWA for Credit Risk: CRE36 IRB Approach: Minimum Requirements to Use IRB Approach. Bank for International Settlements. Available online: https://www.bis.org/basel_framework/chapter/CRE/36.htm?inforce=20190101&export=pdf&pdfid=0 (accessed on 11 March 2019).

Beerbaum, Dirk. 2015. Significant increase in credit risk according to IFRS 9: Implications for financial institutions. *International Journal of Economics and Management Sciences* 4: 1–3. [CrossRef]

Bijak, Katarzyna, and Lyn C. Thomas. 2018. Underperforming performance measures? A review of measures for loss given default models. *Journal of Risk Model Validation* 12: 1–28. [CrossRef]

Breed, Douw Gerbrand, and Tanja Verster. 2017. The benefits of segmentation: Evidence from a South African bank and other studies. *South African Journal of Science* 113: 1–7. [CrossRef]

Breiman, Leo, Jerome Friedman, Richard A. Olsen, and Charles J. Stone. 1984. *Classification and Regression Trees*. Wadsworth: Pacific Grove.

De Jongh, Pieter Juriaan, Tanja Verster, Elzabe Reynolds, Morne Joubert, and Helgard Raubenheimer. 2017. A critical review of the Basel margin of conservatism requirement in a retail credit context. *International Business & Economics Research Journal* 16: 257–74.

European Banking Authority (EBA). 2016. Consultation Paper EBA/CP/2016/10: Draft Guidelines on Credit Institutions' Credit Risk Management Practices and Accounting for Expected Credit Losses. Available online: https://www.eba.europa.eu/documents/10180/1532063/EBA-CP-2016-10+%28CP+on+Guidelines+on+Accounting+for+Expected+Credit%29.pdf (accessed on 3 May 2017).

European Central Bank. 2018. Proposal on ELBE and LGD in-Default: Tackling Capital Requirements after the Financial Crisis. Available online: https://www.ecb.europa.eu/pub/pdf/scpwps/ecb.wp2165.en.pdf?176589bb4b7b020c3d3faffee9b982cd:No2165/June2018 (accessed on 11 February 2019).

Global Public Policy Committee (GPPC). 2016. The Implementation of IFRS 9 Impairment Requirements by Banks: Considerations for Those Charged with Governance of Systemically Important Banks. Global Public Policy Committee. Available online: http://www.ey.com/Publication/vwLUAssets/Implementation_of_IFRS_9_impairment_requirements_by_systemically_important_banks/$File/BCM-FIImpair-GPPC-June2016%20int.pdf (accessed on 25 February 2019).

IFRS. 2014. IRFS9 Financial Instruments: Project Summary. Available online: http://www.ifrs.org/Current-Projects/IASB-Projects/Financial-Instruments-A-Replacement-of-IAS-39-Financial-Instruments-Recognitio/Documents/IFRS-9-Project-Summary-July-2014.pdf (accessed on 31 January 2016).

Joubert, Morne, Tanja Verster, and Helgard Raubenheimer. 2018a. Default weighted survival analysis to directly model loss given default. *South African Statistical Journal* 52: 173–202.

Joubert, Morne, Tanja Verster, and Helgard Raubenheimer. 2018b. Making use of survival analysis to indirectly model loss given default. *Orion* 34: 107–32. [CrossRef]

Kuchibhatla, Maragatha, and Gerda G. Fillenbaum. 2003. Comparison of methods for analyzing longitudinal binary outcomes: Cognitive status as an example. *Aging & Mental Health* 7: 462–68.

Lund, Bruce, and Steven Raimi. 2012. Collapsing Levels of Predictor Variables for Logistic Regression and Weight of Evidence Coding. MWSUG 2012: Proceedings, Paper SA-03. Available online: http://www.mwsug.org/proceedings/2012/SA/MWSUG-2012-SA03.pdf (accessed on 9 April 2019).

Miu, Peter, and Bogie Ozdemir. 2017. Adapting the Basel II advanced internal ratings-based models for International Financial Reporting Standard 9. *Journal of Credit Risk* 13: 53–83. [CrossRef]

Neter, John, Michael H. Kutner, Christopher J. Nachtsheim, and William Wasserman. 1996. *Applied Linear Statistical Models*, 4th ed. WCB McGraw-Hill: New York.

Nguyen, Hoang-Vu, Emmanuel Müller, Jilles Vreeken, and Klemens Böhm. 2014. Unsupervised interaction-preserving discretization of multivariate data. *Data Mining Knowledge Discovery* 28: 1366–97. [CrossRef]

PWC. 2017. IFRS 9 for Banks: Illustrative Disclosures. February. Available online: https://www.pwc.com/ee/et/home/majaastaaruanded/Illustrative_discloser_IFRS_9_for_Banks.pdf (accessed on 8 April 2019).

SAS Institute. 2010. *Predictive Modelling Using Logistic Regression.*. Cary: SAS Institute Inc., Available online: http://support.sas.com/documentation/cdl/en/prochp/67530/HTML/default/viewer.htm#prochp_hpbin_overview.htm (accessed on 6 September 2017).

SAS Institute. 2017. *Development of Credit Scoring Applications Using SAS Enterprise Miner (SAS Course Notes: LWCSEM42)*. Cary: SAS Institute, ISBN 978-1-63526-092-2.

Sheu, Ching-fan. 2000. Regression analysis of correlated binary outcomes. *Behavior Research Methods, Instruments & Computers* 32: 269–73.

Siddiqi, Naeem. 2006. *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*. Hoboken: John Wiley & Sons.

Siddiqi, Naeem. 2017. *Intelligent Credit Scoring: Building and Implementing Better Credit Risk Scorecards*. Hoboken: John Wiley & Sons.

Tevet, Dan. 2013. Exploring model lift: is your model worth implementing? *Actuarial Review* 40: 10–13.

Thomas, Lyn C. 2009. *Consumer Credit Models: Pricing, Profit and Portfolios*. Oxford: Oxford University Press.

Van Berkel, Anthony, and Naeem Siddiqi. 2012. *Building Loss Given Default Scorecard using Weight of Evidence Bins in SAS®Enterprise Miner™*. SAS Institute Inc Paper 141–2012. Cary: SAS Institute.

Verster, Tanja. 2018. Autobin: A predictive approach towards automatic binning using data splitting. *South African Statistical Journal* 52: 139–55.

Volarević, Hrvoje, and Mario Varović. 2018. Internal model for ifrs 9-expected credit losses calculation. *Ekonomski pregled: Mjesečnik Hrvatskog Društva Ekonomista Zagreb* 69: 269. [CrossRef]