mathematics

Article

A Proposed Simulation Technique for Population Stability Testing in Credit Risk Scorecards

Johan du Pisanie, **James Samuel Allison** and **Jaco Visagie** *

School of Mathematical and Statistical Sciences, North-West University, Potchefstroom 2531, South Africa
* Correspondence: jaco.visagie@nwu.ac.za

Abstract: Credit risk scorecards are logistic regression models, fitted to large and complex data sets, employed by the financial industry to model the probability of default of potential customers. In order to ensure that a scorecard remains a representative model of the population, one tests the hypothesis of population stability; specifying that the distribution of customers' attributes remains constant over time. Simulating realistic data sets for this purpose is nontrivial, as these data sets are multivariate and contain intricate dependencies. The simulation of these data sets are of practical interest for both practitioners and for researchers; practitioners may wish to consider the effect that a specified change in the properties of the data has on the scorecard and its usefulness from a business perspective, while researchers may wish to test a newly developed technique in credit scoring. We propose a simulation technique based on the specification of bad ratios, this is explained below. Practitioners can generally not be expected to provide realistic parameter values for a scorecard; these models are simply too complex and contain too many parameters to make such a specification viable. However, practitioners can often confidently specify the bad ratio associated with two different levels of a specific attribute. That is, practitioners are often comfortable with making statements such as "on average a new customer is 1.5 times as likely to default as an existing customer with similar attributes". We propose a method which can be used to obtain parameter values for a scorecard based on specified bad ratios. The proposed technique is demonstrated using a realistic example, and we show that the simulated data sets adhere closely to the specified bad ratios. The paper provides a link to a Github project with the R code used to generate the results.

Keywords: credit risk scorecards; hypothesis testing; population stability; simulation

MSC: 62D99; 62P20

check for updates

Citation: du Pisanie, J.; Allison, J.S.; Visagie, J. A Proposed Simulation Technique for Population Stability Testing in Credit Risk Scorecards. *Mathematics* **2023**, *11*, 492. https://doi.org/10.3390/math11020492

Academic Editors: Niansheng Tang and Shen-Ming Lee

Received: 30 November 2022
Revised: 10 January 2023
Accepted: 12 January 2023
Published: 16 January 2023

1. Introduction and Motivation

Credit scoring is an important technique used in many financial institutions in order to model the probability of default, or some other event of interest, of a potential client. For example, a bank typically has access to data sets containing information pertinent to credit risk, which may be used in order to assess the credit worthiness of potential clients. The characteristics or covariates recorded in such a data set are referred to as attributes throughout; these include information such as income, the total amount of outstanding debt held and the number of recent credit enquiries. A bank may use logistic regression to model an applicant's probability of default as a function of their recorded attributes; these logistic regression models are referred to as credit risk scorecards. In addition to informing the decision as to whether or not a potential borrower is provided with credit, the scorecard is typically used to determine the quoted interest rate. For a detailed treatment of scorecards, see [1] as well as [2].

The development of credit risk scorecards are expensive and time consuming. As a result, once properly trained and validated, a bank may wish to keep a scorecard in use for an extended period, provided that the model continues to be a realistic representation of the attributes of the applicants in the population. One way to determine whether or not a scorecard remains a representative model is to test the hypothesis of population stability. This hypothesis states that the distribution of the attributes remains unchanged over time (i.e., that the distribution of the attributes at present is the same as the distribution observed when the scorecard was developed). When the distribution of the attributes changes, it provides the business with an early indication that the scorecard may no longer be a useful model. Further explanations and examples regarding population stability testing can be found in [3,4] as well as [5].

In the context of testing for population stability, performing scenario testing requires the ability to simulate realistic data sets. To this end, this paper proposes a simple technique for the simulation of such data sets. This enables practitioners to consider scenarios with predefined deviations from specified distributions for the attributes, which allows them to gauge the effects that changes in the distribution of one or more attributes have on the predictions made using the model. Furthermore, the business may also wish to consider the effects of a certain strategy before said strategy is implemented. As a concrete example, consider the case where a bank markets more aggressively to younger people. In this case, they may wish to test the effect of a shift in the distribution of the age of their clients.

The concept of population stability can be further illustrated by means of a simple example. Consider a model that predicts whether someone is wealthy based on a single attribute; the value of the property owned. If this attribute exceeds a specified value, the model predicts that a person is wealthy. Due to house price inflation, the overall prices of houses rise over time. Thus, after a substantial amount of time has passed, the data can no longer be interpreted in the same way as before, and the hypothesis of population stability is rejected, meaning that a new model (or perhaps just a new cut off point) is required.

Population stability metrics measure the magnitude of the change in the distribution of the attributes over time. A number of techniques have been described in the literature, whereby population stability may be tested; see [6,7] as well as [8]. For practical implementations of techniques for credit risk scorecards, see [9] in the statistical software R as well as [10] in Statistical Analysis Software (SAS). The mentioned papers typically provide one or more numerical examples illustrating the use of the proposed techniques. The data sets upon which these techniques are used are typically protected by regulations, meaning that including examples based on the observed data is problematic. As a result, authors often use simulated data. However, the settings wherein these examples are to be found are often oversimplified, stylized and not entirely realistic. This can, at least in part, be ascribed to the difficulties associated with the simulation of realistic data sets. These difficulties arise as a result of the complexity of the nature of the relationship between the attributes and the response.

The data sets typically used for scorecard development have a number of features in common. They are usually relatively large; typical ranges for the number of observations range from one thousand observations to one hundred thousand, while a sample size of one million observations is not unheard of. The data used are multivariate; the number of attributes used varies according to the type of scorecard, what the scorecard will be used for and other factors, but scorecards based on five to fifteen attributes are common. The inclusion of attributes in a scorecard depends on the predictive power of the attribute as well as more practical considerations. These can include the ability to obtain the required data in the future (for example, changing legislation may, in the future, prohibit the inclusion of certain attributes such as gender into the model) as well as the stability of the attribute over the expected lifetime of the scorecard. Care is usually taken to include only attributes with a low level of association with each other so as to avoid the problems associated with multicolinearity.

This paper proposes a simple simulation technique, which may be used for the construction of realistic data sets for use in credit risk scorecards. These data sets contain the attributes of hypothetical customers as well as the associated outcomes. The constructed data sets can be used to perform empirical investigations into the effects of changes in the distribution of the attributes as well as changes in the relationship between these attributes and the outcome. In summary, the advantages of the newly proposed simulation technique are:

1. It is a simple technique.
2. It allows the generation of realistic data sets.
3. These data sets can be used to perform scenario testing.

It should be noted at the outset that the proposed technique is not restricted to the context of credit scoring, or even to the case of logistic regression, but rather has a large number of other modeling applications. However, we restrict our attention to this important special case for the remainder of the paper.

The idea underlying the proposed simulation technique can be summarized as follows. When building a scorecard, practitioners cannot be expected to specify realistic values for the parameters in the model which will ultimately be used. The large number of parameters in the model coupled with the complex relationships between these parameters conspire to make this task almost impossible. However, practitioners can readily be called on to have intuition regarding the bad ratios associated with different states of an attribute. That is, practitioners are often comfortable making statements such as “on average new customers are 1.5 times as likely to default as existing customers with similar attributes”. It should be noted that techniques such as the so-called Delphi method can be used in order to make statement such as these; for a recent reference, see [11].

This paper proposes a technique that can be used to choose parameter values that mimic these specified bad ratios. The inputs required for the proposed technique are the overall bad rate, the specified bad ratios and the marginal distributions of the attributes. It should be noted that the proposed technique can be used to generate data without reference to an existing data set. As such, it is not a data augmentation technique. However, in the event that a reference data set is available, these techniques can be implemented in order to achieve similar goals. An example of a data augmentation technique that can be implemented in this context is so-called generalised adversarial networks, see [12]. Another useful reference on data augmentation is [13]. We emphasize that the newly proposed method can be used in cases where classical data augmentation techniques are not appropriate as the new technique does not require the availability of a data set in order to perform a simulation. As a result, classical data augmentation techniques are not considered further in this paper.

A final noteworthy advantage of the newly proposed technique is its simplicity. Since not all users of scorecards are trained in statistics, the simple nature of the proposed simulation technique (i.e., specifying bad ratios and choosing parameters accordingly) is advantageous.

The remainder of the paper is structured as follows. Section 2 shows several examples of settings in which logistic regression is used in order to model the likelihood of an outcome based on attributes. Here, we demonstrate the need for the proposed simulation procedure. A realistic setting is specified in this section which is used throughout the paper. Section 3 proposes a method that may be used to translate specified bad ratios into model parameters emulating these bad ratios using simulation, followed by parameter estimation. We discuss the numerical results obtained using the proposed simulation technique in Section 4. Section 5 provides some conclusions as well as directions for future research.

2. Motivating Examples

This section outlines several examples. We begin by considering a simple model and we show that the parameters corresponding to a single specified bad ratio can be calculated explicitly, negating the need for the proposed simulation technique. Thereafter, we consider slightly more complicated settings and demonstrate that, in general, no solution exists for a specified set of bad ratios. We also highlight the difficulties encountered when attempting to find the required parameters, should a solution exist. Finally, we consider a realistic model, similar to what one would use in practice.

It should be noted that we consider both discrete and continuous attributes below. There does not seem to be general consensus between practitioners on whether or not continuous attributes should be included in the model, as these attributes are often discretized during the modeling process (some practitioners may argue that we only need consider discrete attributes while others argue against this discretization); for a discussion, see pages 45 to 56 of [1]. Since the number of attributes considered simultaneously using the proposed simulation technique is arbitrary, we may simply chose to replace any continuous attribute by its discretized counterpart. As a result, the techniques described below are applicable in either setting mentioned above.

2.1. A Simple Example

Let X_j be a single attribute, associated with the jth applicant, with two levels, 0 and 1. Denote the respective frequencies with which these values occur by p and $1-p$, respectively;

$$X_j = \begin{cases} 1, \text{ with probability } p, \\ 0, \text{ with probability } 1-p, \end{cases}$$

for $j \in \{1, \ldots, n\}$. Let Y_j be the indicator of default for the jth applicant. Denote the overall bad rate by d; meaning that the unconditional probability of default is $d := P(Y = 1)$. Let γ be the bad ratio of $X_j = 1$ relative to $X_j = 0$. That is, γ is the ratio of the conditional probabilities that $Y_j = 1$ given $X_j = 1$ and $X_j = 0$, respectively; $\gamma := P(Y_j = 1|X_j = 1)/P(Y_j = 1|X_j = 0)$. We may call upon a practitioner to specify appropriate values for d and γ.

Using the information above, we are able to calculate the conditional default rates $d_0 := P(Y_j = 1|X_j = 0)$ and $d_1 := P(Y_j = 1|X_j = 1)$. Simple calculations yield

$$d_0 = \frac{d}{p\gamma + 1 - p}, \quad d_1 = \frac{d\gamma}{p\gamma + 1 - p}.$$

In this setting, building a scorecard requires that the following logistic regression model be fitted:

$$\log\left(\frac{d_j}{1-d_j}\right) = \beta_0 + j\beta_1, \quad j \in \{0, 1\}. \tag{1}$$

Calculating the parameters of the model that give rise to the specified bad ratio requires solving the two equations in (1) in two unknowns. The required solution is calculated to be

$$\beta_0 = \log\left(\frac{d_0}{1-d_0}\right), \quad \beta_1 = \log\left(\frac{d_1}{1-d_1}\right) - \beta_0.$$

As a result, given the values of p, d and γ, we can find a model that perfectly mimics the specified overall probability of default as well as the bad ratio. However, the above example is clearly unrealistically simple.

2.2. Slightly More Complicated Settings

Consider the case where we have three discrete attributes, each with five nominal levels. In this case, the practitioner in question would be required to specify bad ratios for each level of each attribute. This would translate into fifteen equations in fifteen unknowns (since the model would require fifteen parameters in this setting). Solving such a system of equations is already a taxing task, but two points should be emphasized. First, the models used in practice typically have substantially more parameters than fifteen, making the proposition of finding an analytical solution very difficult. Second, there is no guarantee that a solution will exist in this case.

Next, consider the case where a single continuous attribute, say income, is used in the model. When the scorecard is developed, it is common practice to discretize continuous variables such as income into a number of so-called buckets. As a result, the practitioner may suggest, for example, that the population be split into four categories and they may specify a bad ratio for each of these buckets. However, the "true" model underlying the data generates income from a continuous distribution and assigns a single parameter to this attribute in the model. Therefore, this example results in a model with a single parameter which needs to be chosen to satisfy four different constraints (in the form of specified bad ratios). Algebraically, this results in an over specified system in which the number of equations exceed the number of unknowns. In general, an over-specified system of equations cannot be solved.

The two examples above illustrate that, even in unrealistically simple cases, we may not be able to obtain parameters that result in the specified bad ratios.

2.3. A Realistic Setting

We now turn our attention to a realistic setting. Consider the case where ten attributes are used; some of which are continuous while others are discrete. For the discrete case, we distinguish between attributes measured on a nominal scale and attributes measured on a ratio scale. An example of an attribute measured on a nominal scale is the application method used by the applicant as the numerical value assigned to this attribute does not allow direct interpretation. On the other hand, the number of credit cards that an applicant has with other credit providers is measured on an ratio scale, and the numerical value of this attribute allows direct interpretation. In the model used, we treat discrete attributes measured on a ratio scale in the same way as continuous variables; that is, each of these attributes are associated with a single parameter in the model.

As mentioned above, we consider a model containing ten attributes. However, since several discrete attributes are measured on a nominal scale, the number of parameters in the model exceeds the number of attributes. To be precise, let l denote the number of parameters in the model and let m denote the number of attributes measured. Note that $l \geq m$, with equality holding only if no discrete attributes measured on a nominal scale are present. Let $\mathbf{Z}_j = \{Z_{j,1}, \ldots, Z_{j,l}\}$ be the set of attributes associated with the jth applicant. This vector contains the values of observed continuous and discrete, ratio scaled, and attributes. Additionally, $\mathbf{Z}_j$ includes dummy variables capturing the information contained in the discrete, nominal scaled, attributes. Define $\pi_j = E[Y_j|\mathbf{Z}_j]$; the conditional probability of default associated with the jth applicant. The model used can be expressed as

$$\log\left(\frac{\pi_j}{1-\pi_j}\right) = \mathbf{Z}_j{}^\top \boldsymbol{\beta}, \tag{2}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_l)^\top$ is a vector of l parameters.

The names of the attributes included in the model, as well as the scales on which these attributes are measured can be found in Table 1. Care has been taken to use attributes which are often included in credit risk scorecards so as to provide a realistic example. For a discussion of the selection of attributes, see pages 60 to 63 of [1]. Additionally, Table 1 reports the information value of each attribute; this value measures the ability of a specified attribute to predict the value of the default indicator (higher information values indicate higher levels of predictive ability). Consider a discrete attribute with k levels. Let D be the number of defaults in the data set, let D_j be the number of defaults associated with the jth level of this attribute and let n_j be the total number of observations associated with the jth level of this attribute. In this case, the information value of the variable in question is

$$\text{IV} = \sum_{j=1}^{k}\left(\frac{n_j - D_j}{n-D} - \frac{D_j}{D}\right)\log\left(\frac{D(n_j - D_j)}{D_j(n-D)}\right).$$

All calculations below are performed in the statistical software R; see [14].

Table 1. The name, measurement scale and information value of the attributes included in the model.

Name	Scale	Information Value
Gender	Ordinal	0.499
Existing customer	Ordinal	0.441
Number of enquiries	Ratio	0.394
Credit cards with other providers	Ratio	0.515
Province of residence	Ordinal	0.284
Application method	Ordinal	0.222
Age	Ratio	0.164
Total amount outstanding	Ratio	0.083
Income	Ratio	0.182
Balance of recent defaults	Ratio	0.192

For the sake of brevity, we only discuss four of the attributes in detail in the main text of the paper. However, the details of the remaining six attributes, including the numerical results obtained, can be found in Appendix A.

We specify the distribution of the attributes below. For each attribute, we also specify the levels used as well as the bad ratio associated with each of these levels. Care has been taken to use realistic distributions and bad ratios in this example. Admittedly, the process of specifying bad rates is subjective, but we base these values on many years of practical experience in credit scoring, and we believe that most risk practitioners will consider the chosen values plausible. However, it should be stressed that the modeler is not bound to the specific example used here; the proposed technique is general, and the number and distributions of attributes are easily changed. The attributes are treated separately below.

2.4. Existing Customer

Existing customers are usually assumed to be associated with lower levels of risk than is the case for applicants who are not existing customers. This can be due to the fact that existing customers have already shown their ability to repay credit extended to them in the past, or are more likely to pay the company where they have other products. We specify that 80% of applicants are exiting customers and that the bad ratio is 2.7, meaning that the probability of default for a new customer is, on average, 2.7 times higher than the probability of default of an existing customer with the same remaining attributes.

2.5. Credit Cards with Other Providers

This attribute is an indication of the clients exposure to potential credit. A client could, for example, have a low outstanding balance, but through multiple credit cards have access to a large amount of credit. Depending on the type of product being assessed, this could signal higher risk. Table 2 shows the assumed distribution of this attribute together with the specified bad ratios.

Table 2. Credit cards with other providers.

Group	Description	Proportion	Bad Ratio
0	No credit cards at another provider	50%	1.0
1	Credit card at another provider	30%	1.2
2	Credit cards at another provider	15%	1.7
3	Three or more credit cards at another provider	5%	2.5

2.6. Application Method

The method of application is often found to be a very predictive indicator in credit scorecards. A customer actively seeking credit, especially in the unsecured credit space, is often found to be of a higher risk than customers opting in for credit through an outbound method like a marketing call. We distinguish four different application methods:

- Branch—Applications done in the branch.
- Online—Application done through an online application channel.
- Phone—Applications done through a non-direct channel.
- Marketing call—Application done after being prompted by the credit provider.

Table 3 specifies the distribution of this attribute as well as the associated bad ratios.

Table 3. Application method.

Group	Description	Proportion	Bad Ratio
0	Branch	30%	1.0
1	Online	40%	0.5
2	Phone	15%	1.5
3	Marketing Call	15%	0.4

2.7. Age

Younger applicants tend to be higher risk, with risk decreasing as the applicants become older. We assume that the ages of applicants are uniformly distributed between 18 and 75 years. We divide these ages into seven groups, see Table 4.

Table 4. Age.

Group	Proportion	Bad Ratio
18–21	5%	1.00
22–25	7%	0.85
26–30	9%	0.78
31–45	26%	0.66
46–57	21%	0.50
58–63	11%	0.43
64–75	21%	0.31

As was mentioned above, the remaining attributes are discussed in Appendix A. In the next section, we turn our attention to the proposed simulation technique.

3. Proposed Simulation Technique

Having described the details of the attributes included in the model, we turn our attention to finding a model that results in bad ratios approximately equal to those specified. This is done by simulating a large data set, containing attributes as well as default indicators. Thereafter, the parameters of the scorecard are estimated by fitting a logistic regression model to the simulated data. We demonstrate in Section 4 that the resulting parameters constitute a model that closely corresponds to the specified bad ratios and other characteristics. The steps used to arrive at the parameters for the model as well as, ultimately, a simulated data set are as follows:

1. Specify the global parameters.
2. Simulate each attribute separately.
3. Combine the simulated attributes.
4. Fit a logistic regression model.
5. Simulate the final default indicators.

It should be noted that the procedure detailed below assumes independence between the attributes. We opt to incorporate this assumption because it is often made in credit scoring in practice. However, augmenting the procedure below to incorporate dependence between attributes is a simple matter. For example, we can drop the assumption of independent attributes by simulating a group of attributes from a specified copula. Although we do not pursue the use of copulas further below, the reader is referred to [15] for more details.

3.1. Specify the Global Parameters

We specify a fixed, large sample size. It is important that the initial simulated data set be large even in the case where the final simulated sample may be of more modest size, as this will reduce the effect of sample variability. We also specify the overall bad rate. It should be noted that overly small bad rates will tend to decrease the information value of the attributes included in the model (for fixed sets of bad ratios). This is due to the difficulty associated with predicting extremely rare events. We use a sample size of 50,000 and an overall bad rate of 10% to obtain the numerical results shown in the next section.

3.2. Simulate Each Attribute Separately

The next step entails specifying the marginal distribution as well as the bad ratio associated with each attribute. In the case of discrete attributes, a bad ratio is specified for each of the levels of the attribute. In the case of continuous attributes, the attributes are required to be discretized and a bad ratio is specified for each level of the resulting discrete attribute. Given the marginal distribution and the bad ratios of an attribute, we explicitly calculate the bad rate for each level of the attribute. Consider an attribute with k levels and let δ_j be the average bad rate associated with the jth level of the attribute for $j \in \{1, \dots, k\}$. In this case,

$$\delta_j = \frac{\mu_j d}{\sum_{l=1}^{k} \mu_l p_l}, \text{ where } \mu_j = \frac{\gamma_j p_j}{\sum_{l=1}^{k} \gamma_l}.$$

We now simulate a sample of attributes from the specified marginal distribution. Given the values of these attributes, we simulate default indicators from the conditional distribution of these indicators. That is, given that the jth level of the specific attribute is observed, simulate a 1 for the default indicator with probability δ_j.

3.3. Combine the Simulated Attributes

Upon completion of the previous step, we have a realized sample for each of the attributes with a corresponding default indicator. Denoting the sample size by n, the expected number of defaults for each attribute is nd. However, due to sample variation, the number of defaults simulated for the various attributes will differ, which complicates the process of combining the attributes to form a set of simulated attributes for a (simulated) applicant. In order to overcome this problem, we need to ensure that the number of defaults per attribute are equal.

For each attribute, the number of defaults follows a binomial distribution with parameters n and d. As a result, the number of defaults have an expected value nd and variance $nd(1-d)$. Therefore, for large values of n, the ratio of the expected and simulated number of defaults converges to 1 in probability. To illustrate the effect of sample variation, consider the following example. If a sample size of $n = 10^6$ is used and the overall default rate is set to 5%, then the expected number of defaults is 50,000 for each attribute. Due to sample variation, the number of defaults will vary. However, this variation is small when compared to the expected number of defaults; in fact, a 95% confidence interval for the number of defaults is given by $[49\,572; 50\,428]$. Stated differently, the probability that the simulated number of defaults will be within 1% of the expected number is approximately 97.8% in this case, while the probability that the realized number of defaults differ from the expected number by more than 2% is less than 1 in 200,000.

The examples above indicate that the simulated number of defaults will generally be close to nd, and we may assume that changing the simulated number of defaults to exactly nd will not have a large effect on the relationships between the values of the attribute and the default indicator. As a result, we proceed as follows. If the number of defaults exceed nd, we arbitrarily replace 1s with 0s in the default indicator in order to reduce the simulated number of defaults to nd. Similarly, if the number of defaults is less than nd, we replace 0s with 1s.

Following the previous step, the number of defaults per attribute are equal, and we simply combine these attributes according to the default indicator. That is, in order to arrive at the details of a simulated applicant who defaults, we arbitrarily choose one realization of each attributed that resulted in default. The same procedure is used to combine the attributes of applicants who do not default.

3.4. Fit a Logistic Regression Model

We now have a (large) data set containing all of the required attributes as well as the simulated default indicators. We fit a logistic regression model to this data in order to find a parameter set that mimics the specified bad ratios. That is, we estimate the set of regression coefficients in (2). The required estimation is standard, and the majority of statistical analysis packages includes a function to perform the required estimation; the results shown below are obtained using the *glm* function in the *Stats* package of R.

3.5. Simulate the Final Default Indicators

When considering the data set constructed up to this point, the simulated values for the individual attributes are realized from the marginal distribution specified for that attribute. As a result, we need only concern ourselves with the distribution of the default indicator. We now replace the initial default indicator by an indicator simulated from the conditional distribution given the attributes (which is a simple matter since the required parameter estimates are now available). The simulated values of the attributes together with this default indicator constitute the final data set.

The following link contains the R code used for the simulation of a data set using the proposed method; https://bit.ly/3FFLSpp. We emphasize that the user is not bound by the specifications chosen in this paper, as the code is easily amended in order to change the distributions of attributes, to specify other bad ratios and to add or remove attributes from the data set.

4. Performance of the Fitted Model

In order to illustrate the techniques advocated for above, we use the proposed technique to simulate a number of data sets using the specifications in Section 3. Below, we report the means (denoted "Observed bad rate") and standard deviations (denoted "Std dev of obs bad rate") of the observed bad ratios obtained when generating 10,000 data sets, each of size 50,000.

In Tables 5–8, we consider each of the four attributes discussed in the previous section in the main text, while the results associated with the remaining attributes are considered in Appendix B. Tables 5–8 indicate that the average observed bad ratios are remarkably close to the nominally specified bad ratios. Furthermore, the standard deviations of the observed bad ratios are also shown to be quite small, indicating that the proposed method results in data sets in which the specifications provided in Section 3 are closely adhered to.

Table 5. Existing customers.

Group	Description	Specified Bad Rate	Observed Bad Rate	Std Dev of Obs Bad Rate
0	Yes	7.46%	7.48%	0.14%
1	No	20.15%	20.09%	0.46%

Table 6. Credit cards with other providers.

Group	Description	Specified Bad Rate	Observed Bad Rate	Std Dev of Obs Bad Rate
0	No Credit Cards	4.00%	4.70%	0.16%
1	One Credit Card	12.00%	10.43%	0.24%
2	Two Credit Cards	20.00%	19.49%	0.46%
3	Three or more Credit Cards	28.00%	31.90%	1.01%

Table 7. Application method.

Group	Description	Specified Bad Rate	Observed Bad Rate	Std Dev of Obs Bad Rate
0	Branch	12.74%	12.73%	0.32%
1	Online	6.37%	6.39%	0.21%
2	Phone	19.11%	19.05%	0.55%
3	Marketing Call	5.10%	5.12%	0.34%

Table 8. Age.

Group	Description	Specified Bad Rate	Observed Bad Rate	Std Dev of Obs Bad Rate
0	18–21	17.54%	16.82%	0.77%
1	22–25	14.91%	15.19%	0.63%
2	26–30	13.68%	13.89%	0.53%
3	31–45	11.58%	11.46%	0.27%
4	46–57	8.77%	8.66%	0.28%
5	58–63	7.54%	7.28%	0.37%
6	64–75	5.44%	5.82%	0.26%

The marginal distributions of the attributes are not reported in the tables since the average observed proportions coincide with the specified proportions up to 0.01% in all cases. This result is not unexpected, when taking the large sample sizes used into account.

Although less common in practice, smaller sample sizes occur from time to time. This is usually due to constraints placed on the sampling itself; for example, a high cost associated with sampling or regulatory restrictions. When considering smaller sample sizes, the proposed method can still be used. However, in this case the standard deviations of the observed bad rates are increased.

5. Practical Application

The method described above provides a way to arrive at a parametric model, which can be used for simulation purposes, via specification of bad ratios for each attribute considered. One interesting application of this procedure is to specify a deviation from the distribution of the attributes and default indicator and to simulate a second data set. This deviation may, for instance, be in the form of specifying a change in the marginal distribution associated with one or more attributes. The newly simulated data set can then be analyzed in order to gauge the effect of the change to, for example, the overall credit risk of the population.

In practice, a common metric used to measure the level of population stability is the aptly named population stability index (PSI). The PSI quantifies the discrepancy between the observed proportions per level of a given attribute in two samples. Typically, the first data set is observed when the scorecard is developed (we refer to this data set as the *base* data set) and the second is a more recent sample (referred to as the *test* data set). Letting k be the number of levels of the attributes, the PSI is calculated as follows:

$$PSI = \sum_{j=1}^{k} (T_j - B_j) \log\left(\frac{T_j}{B_j}\right), \tag{3}$$

where T_j and B_j, respectively, represent the proportion of the jth level of the attribute in question in the test and base data sets. The following rule-of-thumb for the interpretation of PSI values is suggested in [1]; a value of less than 0.1 indicates that the population shows no substantial changes, a PSI between 0.1 and 0.25 indicates a small change and a PSI of more than 0.25 indicates a substantial change.

It should be noted that the PSI is closely related to the Kullback–Leibler divergence. Let $\mathbf{B} = (B_1, \ldots, B_k)$ and $\mathbf{T} = (T_1, \ldots, T_k)$. The Kullback–Leibler divergence between the base and test populations is defined to be

$$D(\mathbf{B}, \mathbf{T}) = \sum_{j=1}^{k} B_j \log\left(\frac{B_j}{T_j}\right) = -\sum_{j=1}^{k} B_j \log\left(\frac{T_j}{B_j}\right),$$

see [16] as well as [17]. Note that the Kullback–Leibler divergence is an asymmetric discrepancy measure, meaning that the discrepancy between the base and test populations, $D(\mathbf{B}, \mathbf{T})$, need not equal the discrepancy between the test and base populations, $D(\mathbf{T}, \mathbf{B})$. In order to arrive at a symmetric discrepancy measure, one may simply add $D(\mathbf{T}, \mathbf{B})$ to $D(\mathbf{B}, \mathbf{T})$;

$$\begin{aligned} D(\mathbf{B}, \mathbf{T}) + D(\mathbf{B}, \mathbf{T}) &= -\sum_{j=1}^{k} B_j \log\left(\frac{T_j}{B_j}\right) + \sum_{j=1}^{k} T_j \log\left(\frac{B_j}{T_j}\right) \\ &= \sum_{j=1}^{k} (T_j - B_j) \log\left(\frac{T_j}{B_j}\right), \end{aligned}$$

which equals the PSI between the base and test populations. A further discussion of the Kullback–Leibler divergence can be found in [18].

In order to illustrate the use of the PSI, consider the following setup. A single realization of the base data set is simulated using the marginal distributions and the bad ratios specified in Section 2 and Appendix A. We also simulate a test data set using the same specifications, with only the following changes:

- The proportion of existing customers is changed from 80% to 57%. The new distribution is chosen such as to have a PSI value that is approximately 0.25.
- The distribution for the number of enquiries is changed from (30%, 25%, 20%, 15%, 5%, 5%) to (10%, 10%, 20%, 50%, 5%, 5%).

Following these changes, a test data sets is simulated from the distribution specified above and the resulting PSI is calculated for each attribute. This process is repeated 1000 times in order to arrive at 1000 PSI values for each attribute.

In addition to considering the magnitude in the change of the distribution of the attributes, we are interested in measuring the change in the overall credit risk of the population. In order to achieve this, it is standard practice to divide the applicants into various so-called risk buckets based on their probability of default as calculated by the scorecard. In the example used here, we proceed as follows; at the time when the data for the base data set is collected, the applicants may be segmented into ten risk buckets, each containing 10% of the applicants. That is, the $10\%, 20\%, \ldots, 90\%$ quantiles of the probabilities of default of the base data set are calculated. Then, given the test data set, we calculate the proportions of applicants for whom the calculated probability of default is between the $10(j-1)\%$ and $10j\%$ quantiles of the base data set, for $j \in \{1, 2, \ldots, 10\}$. These proportions are then compared to those of the base data set (which are clearly 10% for each risk bucket) in the same way as the proportions associated with the various levels of the attributes are compared. Table 9 contains the average and standard deviations of the PSI calculated for each of the attributes as well as for the risk buckets.

Table 9. Population stability index.

Attribute	Average PSI	Standard Dev of PSI
Gender	0.0001	0.0002
Existing customer	0.2557	0.0102
Number of enquiries	0.7988	0.0178
Credit cards with other providers	0.0005	0.0004
Province of residence	0.0012	0.0005
Application method	0.0005	0.0004
Age	0.0008	0.0004
Total amount outstanding	0.0011	0.0006
Income	0.0008	0.0004
Balance of recent defaults	0.0005	0.0003
Risk buckets	0.0926	0.0061

When considering the results in Table 9, three observations are in order. First, the PSI values calculated for the risk buckets are less than 0.1, indicating that no substantial change in the distribution of the data is observed. Second, the PSI values for the attribute "existing customer" are, on average, 0.2557. Based on the average PSI, the analyst would typically conclude that the variable is unstable as the calculated average PSI value exceeds the cut-off of 0.25. However, in 27.5% of the simulated test data sets, the PSI was calculated to be less than 0.25. This demonstrates that the proposed simulation technique enables us to perform sensitivity analysis in cases where a change in the distribution of the attributes results in PSI values close to the cut-off value of 0.25. When considering the attribute "Number of enquiries", the PSI indicates that a substantial change has occurred. The PSI values calculated for this attribute has an average of 0.7988 and a standard deviation of 0.0178.

6. Conclusions

We propose a simulation technique that can be used in order to generate data sets for use with credit scoring, and we specifically demonstrated the usefulness of this technique for testing population stability. The proposed technique is based on the simple idea of specifying bad ratios and finding parameters that approximately adhere to the specified bad ratios. Using a realistic example, we demonstrate that the proposed technique is able to mimic the specified bad ratios with a high degree of accuracy.

The proposed simulation method enables one to study the properties of population stability metrics in a systematic manner. This allows for the direct comparison of the various measures commonly used in practice in order to identify the strengths and weaknesses of each; research into this topic is currently underway. The proposed method also simplifies the study of newly proposed tests for population stability. Furthermore, another direction for future research is to generalize the proposed simulation technique to the multivariate case; for instance, in the context of multinomial regression. Finally, future research may include extending the proposed methodology to include dependent attributes using copula models. An example of the use of copulas in the context of credit risk can be found in [19].

Author Contributions: Conceptualization, J.d.P., J.S.A. and J.V.; Methodology, J.d.P., J.S.A. and J.V.; Writing—original draft, J.d.P., J.S.A. and J.V.; Writing—review & editing, J.d.P., J.S.A. and J.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: The work of J.S. Allison and I.J.H. Visagie are based on research supported by the National Research Foundation (NRF). Any opinion, finding and conclusion or recommendation expressed in this material is that of the authors and the NRF does not accept any liability in this regard.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Below, we specify the marginal distributions and the specified bad ratios for the characteristics not discussed in detail in the main text of Section 2. Again, we treat each attribute separately.

Appendix A.1. Gender

We assume that 60% of applicants are female and 40% are male, and we specify the bad ratio of males to females to be 3.

Appendix A.2. Number of Enquiries

The number of enquiries is a measure of the client's appetite for credit. A client with a large credit appetite will apply for a number of loans. The number of enquiries provides a view of both the client's successful and unsuccessful applications. Higher numbers of enquiries are often associated with increased levels of risk. Table A1 specifies the distribution associated with various levels of this attribute.

Table A1. Number of enquiries.

Group	Description	Proportion	Bad Ratio
0	No enquiries	30%	1.0
1	One enquiry	25%	1.3
2	Two enquiries	20%	1.8
3	Three enquiries	15%	1.9
4	Four enquiries	5%	2.1
5	Five or more enquiries	5%	2.7

Appendix A.3. Province of Residence

Some provinces are greater economic hubs, which may result in inhabitants with lower levels of credit risk. Table A2 shows the marginal distribution as well as bad ratios assumed for the 9 provinces of South Africa.

Table A2. Province of residence.

Group	Description	Proportion	Bad Ratio
0	Gauteng	40%	1.0
1	Western Cape	30%	0.7
2	KwaZulu Natal	7%	1.8
3	Mpumalanga	5%	1.5
4	North West	5%	3.0
5	Limpopo	4%	2.5
6	Eastern Cape	4%	2.0
7	Northern Cape	3%	4.0
8	Free State	2%	1.2

Appendix A.4. Total Amount Outstanding

An applicant's total amount outstanding is an indication of the current indebtedness and provides a view of the client's previous commitments. Excessively low or high levels of this variable may be associated with higher levels of risk; i.e., a customer with no outstanding amount could be a result of not being able to obtain credit while very high levels of this attribute may indicate difficulty in paying current commitments. The marginal distribution specified for this attribute is standard lognormal, rescaled by a factor of 10,000. The lognormal distribution is chosen since its shape is reminiscent of the empirical distribution typically observed in practice, while the scaling factor is incorporated in order to ensure that the numbers used are of a realistic magnitude. The resulting proportions and bad ratios can be found in Table A3.

Table A3. Total amount outstanding.

Group	Grouping	Proportion	Bad Ratio
0	0–5000	24.4%	1.0
1	5000–10,000	25.6%	1.2
2	10,000–25,000	32.0%	2.0
3	25,000–100,000	16.9%	2.1
4	more than 100,000	1.1%	0.8

Appendix A.5. Income

Income is a strong indicator of the ability to repay debt and it is often used directly or indirectly in the scoring process. Direct use occurs through inclusion into the scoring model as an attribute, while indirect use can be accomplished through using income as an entry criteria for the application. The distribution used for income is a mixture with several local models. The associated proportions and bad ratios can be found in Table A4.

Table A4. Income.

Group	Grouping	Proportion	Bad Ratio
0	0–5000	3.2%	3.0
1	5000–11,000	15.6%	2.5
2	11,000–20,000	20.4%	2.0
3	20,000–30,000	21.8%	1.4
4	30,000–70,000	24.0%	1.2
5	more than 70,000	15.0%	1.0

Balance of Recent Defaults

Recent defaults are an indication that a customer is no longer able to pay their debts. This attribute specifically speaks to customers that have recently defaulted, as all customers without defaults are grouped at zero. Table A5 specifies a distribution in which the majority of applicants have recent defaults with a value of less than 1000 units, indicating that the majority of applicants have not defaulted recently.

Table A5. Balance of recent defaults.

Group	Grouping	Proportion	Bad Ratio
0	0–1000	60.0%	1.0
1	1000–3000	1.1%	1.1
2	3000–5000	2.1%	2.0
3	5000–30,000	18.9%	2.5
4	30,000–1,000,000	18.0%	3.0
5	more than 1,000,000	0.0%	3.3

Appendix B

Tables A6–A11 report the specified bad rate, the average observed bad rate as well as the standard deviation of this bad rate for each of the attributes not treated in the main text of Section 4.

Table A6. Gender.

Group	Description	Specified Bad Rate	Observed Bad Rate	Std Dev of Obs Bad Rate
0	Female	5.56%	5.58%	0.16%
1	Male	16.67%	16.62%	0.27%

Table A7. Number of enquiries.

Group	Description	Specified Bad Rate	Observed Bad Rate	Std Dev of Obs Bad Rate
0	No Enquiries	6.62%	6.97%	0.23%
1	One Enquiry	8.61%	8.62%	0.25%
2	Two Enquiries	11.92%	10.89%	0.30%
3	Three Enquiries	12.58%	12.74%	0.39%
4	Four Enquiries	13.91%	14.96%	0.73%
5	Five or more Enquiries	17.88%	18.28%	0.84%

Table A8. Province of residence.

Group	Description	Specified Bad Rate	Observed Bad Rate	Std Dev of Obs Bad Rate
0	Gauteng	7.78%	7.79%	0.22%
1	Western Cape	5.45%	5.47%	0.24%
2	KwaZulu Natal	14.01%	14.00%	0.74%
3	Mpumalanga	11.67%	11.67%	0.83%
4	North West	23.35%	23.27%	1.05%
5	Limpopo	19.46%	19.40%	1.11%
6	Eastern Cape	15.56%	15.51%	1.05%
7	Northern Cape	31.13%	30.98%	1.50%
8	Free State	9.34%	9.32%	1.22%

Table A9. Amount outstanding.

Group	Description	Specified Bad Rate	Observed Bad Rate	Std Dev of Obs Bad Rate
0	0–5000	6.43%	8.52%	0.26%
1	5000–10,000	7.72%	9.01%	0.25%
2	10,000–25,000	12.86%	10.78%	0.23%
3	25,000–100,000	13.51%	11.93%	0.36%
4	more than 100,000	5.15%	12.17%	2.98%

Table A10. Income.

Group	Description	Specified Bad Rate	Observed Bad Rate	Std Dev of Obs Bad Rate
0	0–5000	19.07%	14.51%	0.95%
1	5000–11,000	15.89%	12.65%	0.44%
2	11,000–20,000	12.71%	11.66%	0.34%
3	20,000–30,000	8.90%	10.28%	0.29%
4	30,000–70,000	7.63%	9.70%	0.30%
5	more than 70,000	6.36%	4.08%	0.44%

Table A11. Balance of recent defaults.

Group	Description	Specified Bad Rate	Observed Bad Rate	Std Dev of Obs Bad Rate
0	1000–3000	6.11%	8.41%	0.18%
1	3000–5000	6.72%	4.51%	0.91%
2	5000–30,000	12.22%	5.78%	0.75%
3	30,000–1,000,000	15.28%	8.21%	0.32%
4	more than 1,000,000	18.33%	17.99%	0.45%

References

1. Siddiqi, N. *Credit Risk Scorecards*; John Wiley and Sons, Inc.: Hoboken, NJ, USA, 2006.
2. Siddiqi, N. *Intelligent Credit Scoring: Building and Implementing Better Credit Risk Scorecards*; John Wiley and Sons, Inc.: Hoboken, NJ, USA, 2016.

3. Anderson, R. *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*; Oxford University Press: Oxford, UK, 2007.
4. Karakoulas, G. Empirical Validation of Retail Credit-Scoring Models. *RMA J.* **2004**, *87*, 56–60.
5. Lewis, E.M. *An Introduction to Credit Scoring*; Athena Press: London, UK, 1994.
6. Taplin, R.; Hunt, C. The Population Accuracy Index: A New Measure of Population Stability for Model Monitoring. *Risks* **2019**, *7*, 53. [CrossRef]
7. Yurdakul, B.; Naranjo, J. Statistical properties of the population stability index. *J. Risk Model Valid.* **2019**, *14*. [CrossRef]
8. Du Pisanie, J.; Visagie, I.J.H. On testing the hypothesis of population stability for credit risk scorecards. *ORiON J.* **2020**, *36*, 19–34. [CrossRef]
9. Fan, D. creditmodel: Toolkit for Credit Modelling; R Package Version 1.1.9. 2020. Available online: https://cran.r-project.org/web/packages/creditmodel/index.html (accessed on 1 November 2022).
10. Pruitt, R. The Applied Use of Population Stability Index (PSI) in SAS Enterprise Miner Posters. *SAS Global Forum*. 2010. Available online: http://support.sas.com/resources/papers/proceedings10/288-2010.pdf (accessed on 1 November 2022).
11. Markmann, C.; Spickermann, A.; von der Gracht, H.A.; Brem, A. Improving the question formulation in Delphi-like survey: Analysis of the effects of abstract language and amount of information on response behavior. *Futur. Foresight Sci.* **2020**, *3*, e56. [CrossRef]
12. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *Adv. Neural Inf. Process. Syst.* **2014**, *3*. [CrossRef]
13. Bedrick, E.J.; Christensen, R.; Johnson, W. A new perspective on priors for generalized linear models. *J. Am. Stat. Assoc.* **1996**, *91*, 1450–1460. [CrossRef]
14. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2019.
15. Nelson, R.G. *An Introduction to Copulas*; Springer: New York, NY, USA, 2006.
16. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]
17. Kullback, S. *Information Theory and Statistics*; John Wiley and Sons, Inc.: London, UK, 1959.
18. Wu, D.; Olson, D. Enterprise risk management: Coping with model risk in a large bank. *J. Oper. Res. Soc.* **2010**, *61*, 179–190. [CrossRef]
19. Lu, M.J.; Chen, C.Y.H.; H ardle, W.K. Copula-based factor model for credit risk analysis. *Rev. Quant. Financ. Account.* **2017**, *49*, 949–971. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.