Applied Mathematical Sciences, Vol. 8, 2014, no. 65, 3229 - 3242
HIKARI Ltd, www.m-hikari.com
http://dx.doi.org/10.12988/ams.2014.44300

A Necessary Condition for a Good Binning Algorithm in Credit Scoring

Guoping Zeng

Elevate/Think Finance
4150 International Plaza
Fort Worth, TX 76109, USA

Abstract

Binning is a categorization process to transform a continuous variable into a small set of groups or bins. Binning is widely used in credit scoring. In particular, it can be used to define the Weight of Evidence (WOE) transformation. In this paper, we first derive an explicit solution to a logistic regression model with one independent variable that has undergone a WOE transformation. We then use this explicit solution as a necessary condition for a good binning algorithm, thus providing a simple way to catch binning errors.

Keywords: Binning, Weight of Evidence (WOE), Maximum Likelihood Estimate, Logistic Regression, Credit Scoring.

1 Introduction

Binning is a categorization process to transform a continuous variable into a small set of groups or bins. Binning is widely used in credit scoring. While binning can be used to find Kolmogorov-Smirnov (KS) and lift chart from scores, binning is more frequently used at an early stage to select variables in credit scoring. Similar attributes (values) of an independent variable x are grouped into the same bin to enhance the predictive power. After binning is done, Information Value (IV) and other types of metric divergence measures can be calculated [8]. In particular, binning can be used to introduce Weight of Evidence (WOE) transformations for continuous variables.

Four binning algorithms are commonly used in credit scoring: equal-width binning, equal-size binning, optima binning and Multi-Interval Discretization binning in machine learning. In equal-width binning [6], the values of $\boldsymbol{x}$ is divided into a pre-defined number of equal width intervals. In equal-size binning, the attributes are sorted first, and then divided into a pre-defined number of equal-size bins. If $\boldsymbol{x}$ has distinct values, all the bins will have the same number of observations except the last one which may have fewer observations. In reality, $\boldsymbol{x}$ may have repeating values. In this case, the repeating attributes must stay in the same bin. In SAS, PROC RANK can be used to do equal-size binning [8]. Specifically, PROC RANK computes the ranks of the values, uses GROUPS option to specify the number of bins, and handles ties of values. In optimal binning [6], $\boldsymbol{x}$ is divided into a large number of initial equal-width bins, say 50. These bins are then treated as categories of a nominal variable and grouped to the required number of segments in a tree structure. Multi-Interval Discretization binning [3] is the entropy minimization for binary discretizing the range of a continuous variable into multiple intervals, and recursively define the best bins.

A good binning algorithm should follow the following guidelines [7]:

- Missing values are binned separately.
- Each bin should contain at least 5% of observations.
- No bins have 0 accounts for good or bad.

WOE is a quantitative method for combining evidence in support of a statistical hypothesis [4]. WOE is widely used in credit scoring to separate good accounts and bad accounts. It compares the proportion of good accounts to bad accounts at each attribute level, and measures the strength of the attributes of an independent variable in separating good and bad accounts.

In this paper, we first derive an explicit solution to a logistic regression model with one independent variable that has undergone the WOE transformation. We then use this explicit solution as a necessary condition, thus providing a simply way to catch binning errors.

The rest of the paper is organized as follows. In Section 2, the basic of logistic regression and maximum likelihood estimate are reviewed. In Section 3, we derive an explicit solution to a logistic regression model with one continuous variable that has undergone the WOE transformation. Section 4 states the necessary condition for good binning and presents a numerical example to catch binning errors. The paper is concluded in Section 5.

2 Logistic Regressions and Maximum Likelihood Estimate

To start with, let's assume that $x = (x_1, x_2, \dots, x_p)$ are the vector of p independent variables and y is the dichotomous dependent variable. Assume we

have a sample of N independent observations $(x_{i1}, x_{i2}, \dots, x_{ip}, y_i), i = 1, 2, \dots, N,$ where y_i denotes the value of y (0 for Good status and 1 for Bad status) and $x_{i1}, x_{i2}, \dots, x_{ip}$ are the values of $x_1, x_2, \dots, x_p$ for the i-th observation, respectively.

To adopt standard notation in logistic regression [9], we use the quantity $\pi(x) = E(y|x)$ to represent the conditional mean of y given x. The logistic regression model is given by the equation

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}. \tag{2.1}$$

The logit transformation of $\pi(x)$ is

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p. \tag{2.2}$$

The likelihood function for logistic regression can be expressed as follows

$$l(\beta) = \prod_{i=1}^{N} \pi(x_{i1}, x_{i2}, \dots, x_{ip})^{y_i}\left[1 - \pi(x_{i1}, x_{i2}, \dots, x_{ip})\right]^{1-y_i} \tag{2.3}$$

where β is the vector $(\beta_0, \beta_1, \dots \beta_p)$.

Note that if y_i is known, either 0 or 1, the 2 terms in the product of (2.3) reduces to only one term as the other term will have value of 1.

The solution to the maximum likelihood of logistic regression is an estimate of β which maximizes the expression (2.3). Since it is easier to work with the log of equation, the log likelihood is instead used

$$L(\beta) = \ln[l(\beta)] =$$
$$\sum_{i=1}^{N}\{y_i \ln[\pi(x_{i1}, x_{i2}, \dots, x_{ip})] + (1 - y_i)\ln[1 - \pi(x_{i1}, x_{i2}, \dots, x_{ip})]\}. \tag{2.4}$$

The solution β to (2.4) is called the maximum likelihood estimate. The maximum likelihood estimate of $\pi(x_{i1}, x_{i2}, \dots, x_{ip})$ will be denoted by $\hat{\pi}(x_{i1}, x_{i2}, \dots, x_{ip})$ or simply $\hat{\pi}(x_i)$. It follows from (3.1) that $0 < \hat{\pi}(x_i) < 1$.

One well-known approach to maximizing a function is to differentiate it with respect to β, set the derivative to 0, and then solve the resulting equations. Differentiating $L(\beta)$ with respect to β_0 and setting it to 0, one obtains

$$\sum_{i=1}^{N} y_i = \sum_{i=1}^{N} \pi(x_{i1}, x_{i2}, \dots, x_{ip}). \tag{2.5}$$

Differentiating $L(\beta)$ with respect to β_j for $j = 1,2, \dots, p$, and setting it to 0,

$$\sum_{i=1}^{N} \boldsymbol{x_{ij} y_i} = \sum_{i=1}^{N} x_{ij} \pi(x_{i1}, x_{i2}, \dots, x_{ip}). \quad (2.6)$$

Since binning is done for each independent variable, from now on we shall focus on a single independent variable x. In this case, Equations (2.5) and (2.6) become

$$\sum_{i=1}^{N} \boldsymbol{y_i} = \sum_{i=1}^{N} \pi(x_i). \quad (2.7)$$

$$\sum_{i=1}^{N} \boldsymbol{x_i y_i} = \sum_{i=1}^{N} \boldsymbol{x_i} \pi(x_i). \quad (2.8)$$

Remark 2.1. If x has missing values, they will be simply ignored in logistic regression. Specifically, missing values will be taken out from (2.7) and (2.8). Thus, the 4 summations in (2.7) and (2.8) are taken for all non-missing values of x.

Let's study the existence and uniqueness of the maximum likeliness estimate. Since the log likelihood is globally concave, the maximum likeliness estimate can have at most one solution [1]. Therefore, if the maximum likeliness estimate has a solution, it must be unique. However, there are situations in which the log likelihood function has no maximum and hence the maximum likelihood estimate does not exist. For instance, in case of complete separation or quasi-complete separation [2] the maximum likelihood estimate does not exist. Consider a dataset with 20 observations, where y = 0 if x is $-10, -9, \dots, -2, -1$ and 1 if x is $1, 2, \dots, 9, 10$.

The left hand side of (2.8) is 55, and the absolute value of right hand side of (2.8) is

$$\left| \sum_{i=1}^{10} \boldsymbol{i} \left[\frac{e^{\beta_0 + \beta_1 i}}{1 + e^{\beta_0 + \beta_1 i}} - \frac{e^{\beta_0 - \beta_1 i}}{1 + e^{\beta_0 - \beta_1 i}} \right] \right| \leq \sum_{i=1}^{10} \boldsymbol{i} \left| \frac{e^{\beta_0 + \beta_1 i}}{1 + e^{\beta_0 + \beta_1 i}} - \frac{e^{\beta_0 - \beta_1 i}}{1 + e^{\beta_0 - \beta_1 i}} \right|$$
$$< \sum_{i=1}^{10} \boldsymbol{i} = 55.$$

Hence, (2.8) and so the maximum likelihood estimate has no solution.

On the other hand, if the simultaneous equations (2.7) and (2.8) have a solution, a natural question arises: can it be explicitly solved. For some models and data, the answer is yes. For instance, when the system has only one independent variable and this independent variable is dichotomous [5]. For most models, they cannot be explicitly solved and must be solved by numerical methods.

3 An Explicit Solution to Maximum Likelihood Estimate after WOE Transformation

Mathematically, WOE is defined as the logarithm of the ratio of the odds of Bad-to-Good in the attribute level to the odds of Bad-to-Good in the entire sample. Table 3.1 demonstrates the use of WOE for independent variable x. For convenience, we put all missing values to the last group. If x is a categorical variable, each category is a group and has the same value. If x is a continuous variable, it is first binned into groups. Moreover, the values of x are sorted in the increasing order inside each group and across groups so that

$$x_1 \leq x_2 \leq \cdots \leq x_{n_1} < x_{n_1+1} \leq x_{n_1+2} \leq \cdots \leq x_{n_2} < \cdots < x_{n_{k-1}+1} \leq x_{n_{k-1}+2} \leq \ldots \leq x_{n_k}.$$

Group	**x**	**Good Accounts**	**Bad Accounts**	**WOE**
1	x_1 x_2 … x_{n_1}	g_1	b_1	$\ln \frac{b_1/(b_1+b_2 \ldots + b_{k+1})}{g_1/(g_1+g_2 \ldots + g_{k+1})}$
2	x_{n_1+1} x_{n_1+2} … x_{n_2}	g_2	b_2	$\ln \frac{b_2/(b_1+b_2 \ldots + b_{k+1})}{g_2/(g_1+g_2 \ldots + g_{k+1})}$
…	…	…	…	…
k	$x_{n_{k-1}+1}$ $x_{n_{k-1}+2}$ … x_{n_k}	g_k	b_k	$\ln \frac{b_k/(b_1+b_2 \ldots + b_{k+1})}{g_k/(g_1+g_2 \ldots + g_{k+1})}$
$k+1$	x_{n_k+1} x_{n_k+2} … $x_{n_{k+1}}$	g_{k+1}	b_{k+1}	$\ln \frac{b_{k+1}/(b_1+b_2 \ldots + b_{k+1})}{g_{k+1}/(g_1+g_2 \ldots + g_{k+1})}$

Table 3.1: Binning and WOE's

Denote the number of good accounts and number of bad accounts at group j by g_j and b_j, respectively. Then, b_j is equal to the sum of y in group j, that is,

$$b_j = \sum_{i=n_{j-1}+1}^{n_j} y_i.$$

Since the observations with missing y will be ignored, we may assume y has no missing values without loss of generality. Hence, $g_1 + b_1 = n_1$ represents the number of accounts in bin 1, and $g_j + b_j = n_j - n_{j-1}$ the number of accounts at group j for $j = 2, 3, \dots, k+1$. For convenience, we define n_{-1} as 0.

Note that all the values of x in each group have the same WOE. In this case, the values of x are transformed into grouped WOE's, thus reducing complexities to the modeling.

Theorem 3.1. When a logistic regression model is fitted with one independent variable that has undergone a WOE transformation, the maximum likelihood estimate has an explicit solution $\beta_0 = ln\left(\frac{b}{g}\right)$ and $\beta_1 = 1$, where b and g are the number of bad accounts and number of good accounts, respectively.

Proof. From the uniqueness of the maximum likeliness estimate, it is sufficient to verify $\beta_0 = \frac{b}{g}$ and $\beta_1 = 1$ satisfy (2.7) and (2.8) after the WOE transformation, where $b = b_1 + b_2 + \cdots + b_{k+1}$ and $g = g_1 + g_2 + \cdots + g_{k+1}$. There are two cases to consider.

Case I: **No groups have 0 good accounts or 0 bad accounts**, that is, g_i and b_i are all positive for $j = 1, 2, \dots, k+1$.

We first verify (2.7). Note that the new independent variable WOE has the same value inside each group. Substituting $\beta_0 = \frac{b}{g}$ and $\beta_1 = 1$ into the left hand side of (2.7), we obtain

$$\sum_{i=1}^{n_{k+1}} \pi(WOE(x_i))$$

$$= \sum_{i=1}^{n_1} \pi(WOE(x_i)) + \sum_{i=n_1+1}^{n_2} \pi(WOE(x_i)) + \cdots$$

$$+ \sum_{i=n_{k-1}+1}^{n_k} \pi(WOE(x_i)) + \sum_{i=n_k+1}^{n_{k+1}} \pi(WOE(x_i))$$

$$= (g_1 + b_1)\frac{e^{\ln\frac{b}{g}+\ln\frac{\frac{b_1}{b}}{\frac{g_1}{g}}}}{1+e^{\ln\frac{b}{g}+\ln\frac{\frac{b_1}{b}}{\frac{g_1}{g}}}} + (g_2 + b_2)\frac{e^{\ln\frac{b}{g}+\ln\frac{\frac{b_2}{b}}{\frac{g_2}{g}}}}{1+e^{\ln\frac{b}{g}+\ln\frac{\frac{b_2}{b}}{\frac{g_2}{g}}}}$$

$$+ \cdots (g_k + b_k)\frac{e^{\ln\frac{b}{g}+\ln\frac{\frac{b_k}{b}}{\frac{g_k}{g}}}}{1+e^{\ln\frac{b}{g}+\ln\frac{\frac{b_k}{b}}{\frac{g_k}{g}}}} + (g_{k+1} + b_{k+1})\frac{e^{\ln\frac{b}{g}+\ln\frac{\frac{b_{k+1}}{b}}{\frac{g_{k+1}}{g}}}}{1+e^{\ln\frac{b}{g}+\ln\frac{\frac{b_{k+1}}{b}}{\frac{g_{k+1}}{g}}}}$$

$$= \sum_{j=1}^{k+1} (g_j + b_j)\frac{\frac{b}{g}\times\frac{\frac{b_j}{b}}{\frac{g_j}{g}}}{1+\frac{b}{g}\times\frac{\frac{b_j}{b}}{\frac{g_j}{g}}}$$

$$= \sum_{j=1}^{k+1} (g_j + b_j)\frac{\frac{b_j}{g_j}}{1+\frac{b_j}{g_j}}$$

$$= \sum_{j=1}^{k+1} b_j = \sum_{j=1}^{k+1}\left(\sum_{i=n_{j-1}+1}^{n_j} y_i\right) = \sum_{i=1}^{n_{k+1}} y_i.$$

Next, we verify (2.8).

$$\sum_{i=1}^{n_{k+1}} WOE(x_i)\pi(WOE(x_i)) = \sum_{j=1}^{k+1}(g_j + b_j)\ln\frac{\frac{b_j}{b}}{\frac{g_j}{g}} \times \frac{e^{\ln\frac{b}{g}+\ln\frac{\frac{b_j}{b}}{\frac{g_j}{g}}}}{1+e^{\ln\frac{b}{g}+\ln\frac{\frac{b_j}{b}}{\frac{g_j}{g}}}} =$$

$$\sum_{j=1}^{k+1}(g_j + b_j)\frac{\frac{b}{g}\times\frac{\frac{b_j}{b}}{\frac{g_j}{g}}}{1+\frac{b}{g}\times\frac{\frac{b_j}{b}}{\frac{g_j}{g}}} \times \ln\frac{\frac{b_j}{b}}{\frac{g_j}{g}} = \sum_{j=1}^{k+1}(g_j + b_j)\frac{\frac{b_j}{g_j}}{1+\frac{b_j}{g_j}} \times$$

$$\ln\frac{\frac{b_j}{b}}{\frac{g_j}{g}} = \sum_{j=1}^{k+1} b_j \ln\frac{\frac{b_j}{b}}{\frac{g_j}{g}} = \sum_{j=1}^{k+1}(y_{n_{j-1}+1} + y_{n_{j-1}+2} + \cdots + y_{n_j})\ln\frac{\frac{b_j}{b}}{\frac{g_j}{g}} =$$

$$\sum_{j=1}^{k+1}\left(\sum_{i=n_{j-1}+1}^{n_j} y_i\right)\ln\frac{\frac{b_j}{b}}{\frac{g_j}{g}} = \sum_{i=1}^{n_1} y_i \ln\frac{\frac{b_1}{b}}{\frac{g_1}{g}} + \sum_{i=n_1+1}^{n_2} y_i \ln\frac{\frac{b_2}{b}}{\frac{g_2}{g}} + \cdots +$$

$$\sum_{i=n_{k-1}+1}^{n_k} y_i \ln\frac{\frac{b_k}{b}}{\frac{g_k}{g}} + \sum_{i=n_k+1}^{n_{k+1}} y_i \ln\frac{\frac{b_{k+1}}{b}}{\frac{g_{k+1}}{g}} =$$

$$\sum_{i=1}^{n_1} y_i WOE(x_i) + \sum_{i=n_1+1}^{n_2} y_i WOE(x_i) + \cdots + \sum_{i=n_{k-1}+1}^{n_k} y_i WOE(x_i) +$$

$$\sum_{i=n_k+1}^{n_{k+1}} y_i WOE(x_i) =$$

$$\sum_{i=1}^{n_k} y_i WOE(x_i) + \sum_{i=n_k+1}^{n_{k+1}} y_i WOE(x_i) = \sum_{i=1}^{n_{k+1}} y_i WOE(x_i).$$

Note that if x has no missing values, bin $k + 1$ does not exist. The above proof still holds after the last row of table 1 is removed.

Case II: **Some groups have 0 bad accounts or 0 good accounts, that is, $g_i = 0$ or $b_i = 0$** for some i.

In this case, WOEs are not defined and so the new independent variable WOE will have missing values in these groups. Assume among the k + 1 groups only groups $s_1, s_2, \ldots, s_m$ do not have 0 bad account or 0 good account, where $1 \le s_1 < s_2 \ldots < s_m \le k + 1$. Then, the left hand side of (2.7) becomes

$$\sum_{i=n_{s_1-1}+1}^{n_{s_1}} \pi\big(WOE(x_i)\big) + \cdots + \sum_{i=n_{s_m-1}+1}^{n_{s_m}} \pi\big(WOE(x_i)\big)$$

$$= \big(g_{s_1} + b_{s_1}\big)\frac{e^{\ln\frac{b}{g}+\ln\frac{\frac{b_{s_1}}{b}}{\frac{g_{s_1}}{g}}}}{1+e^{\ln\frac{b}{g}+\ln\frac{\frac{b_{s_1}}{b}}{\frac{g_{s_1}}{g}}}} + \cdots + \big(g_{s_m} + b_{s_m}\big)\frac{e^{\ln\frac{b}{g}+\ln\frac{\frac{b_{s_m}}{b}}{\frac{g_{s_m}}{g}}}}{1+e^{\ln\frac{b}{g}+\ln\frac{\frac{b_{s_m}}{b}}{\frac{g_{s_m}}{g}}}}$$

$$= \big(g_{s_1} + b_{s_1}\big)\frac{\frac{b_{s_1}}{g_{s_1}}}{1+\frac{b_{s_1}}{g_{s_1}}} + \cdots + \big(g_{s_m} + b_{s_m}\big)\frac{\frac{b_{s_m}}{g_{s_m}}}{1+\frac{b_{s_m}}{g_{s_m}}}$$

$$= b_{s_1} + b_{s_2} + \cdots + b_{s_m}$$

$$= \sum_{i=n_{s_1-1}+1}^{n_{s_1}} y_i + \cdots + \sum_{i=n_{s_m-1}+1}^{n_{s_m}} y_i.$$

Next, we verify (2.8).

$$\sum_{i=n_{s_1-1}+1}^{n_{s_1}} WOE(x_i)\pi\big(WOE(x_i)\big) + \cdots + \sum_{i=n_{s_m-1}+1}^{n_{s_m}} WOE(x_i)\pi\big(WOE(x_i)\big)$$

$$= \big(g_{s_1} + b_{s_1}\big)\ln\frac{\frac{b_{s_1}}{b}}{\frac{g_{s_1}}{g}} \times \frac{e^{\ln\frac{b}{g}+\ln\frac{\frac{b_{s_1}}{b}}{\frac{g_{s_1}}{g}}}}{1+e^{\ln\frac{b}{g}+\ln\frac{\frac{b_{s_1}}{b}}{\frac{g_{s_1}}{g}}}} + \cdots + \big(g_{s_m} + b_{s_m}\big)\ln\frac{\frac{b_{s_m}}{b}}{\frac{g_{s_m}}{g}} \times \frac{e^{\ln\frac{b}{g}+\ln\frac{\frac{b_{s_m}}{b}}{\frac{g_{s_m}}{g}}}}{1+e^{\ln\frac{b}{g}+\ln\frac{\frac{b_{s_m}}{b}}{\frac{g_{s_m}}{g}}}}$$

$$= b_{s_1}\ln\frac{\frac{b_{s_1}}{b}}{\frac{g_{s_1}}{g}} + \cdots + b_{s_m}\ln\frac{\frac{b_{s_m}}{b}}{\frac{g_{s_m}}{g}}$$

$$= \left(\sum_{i=n_{s_1-1}+1}^{n_{s_1}} y_i\right)\ln\frac{\frac{b_{s_1}}{b}}{\frac{g_{s_1}}{g}} + \cdots \left(\sum_{i=n_{s_m-1}+1}^{n_{s_m}} y_i\right)\ln\frac{\frac{b_{s_m}}{b}}{\frac{g_{s_m}}{g}}$$

$$= \sum_{i=n_{s_1-1}+1}^{n_{s_1}} y_i\, WOE(x_i) + \cdots + \sum_{i=n_{s_m-1}+1}^{n_{s_m}} y_i\, WOE(x_i).$$

Remark 3.1. Refaat [6] uses an example to find an explicit solution to the maximum likelihood estimate for an independent categorical variable. We have extended to an independent continuous variable, considered missing values and bins with 0 good or bad accounts, and analytically verified the solution.

4 A Necessary Condition

Theorems 3.1 can serve as a necessary condition for a good binning algorithm.

Corollary 4.1. A necessary condition for a good binning is that $\beta_0 = ln\left(\frac{b}{g}\right)$ and $\beta_1 = 1$ when a logistic regression model is fitted with one independent variable that has undergone a WOE transformation.

We can use Corollary 4.1 to catch binning errors of a binning algorithm. To do this, we first perform the WOE transformation after binning. Next, we run logistic regression. If the slope is not 1 or the intercept is not $ln\left(\frac{b}{g}\right)$, then this binning algorithm is not good. In practice, we should bear with computational errors. Let's look at a numerical example. We use an imaginary dataset with age as the independent variable and y as the dependent variable.

```
data age1;
    input age y @@;
    datalines;

   10 0 10 0 10 1 10 0 10 1 10 0 10 0 10 0 10 0 10 0
   10 1 10 0 10 0 10 0 10 1 10 1 10 0 10 0 10 0 10 0
   10 0 10 0 10 0 10 0 10 0 10 0 10 0 10 1 10 0 10 0
   10 0 10 0 10 0 10 1 10 0 10 0 10 0 10 0 10 0 10 0
   10 0 10 0 10 0 10 0 10 0 10 0 10 0 10 1 10 0 10 1
   20 0 20 0 20 0 20 1 20 0 20 0 20 0 20 1 20 0 20 0 20 0
   20 1 20 0 20 0 20 0 20 0 20 0 20 1 20 0 20 1 20 0
   20 0 20 0 20 0 20 1 20 0 20 0 20 0 20 0 20 0 20 0
   30 1 30 0 30 0 30 1 30 1 30 0 30 0 30 0 30 0 30 0
   40 0 40 0 40 0 40 0 40 0 40 0 40 0 40 0 40 0 48 0
   . 0 . 0 . 1 . 0 . 0 . 0 . 1 . 0 . 0 . 0 . 1
   ;
run;
```

The data after the datalines statement in the above SAS program represent the values of age and y in turn. Note that age has some missing values. Using equal-size binning with SAS PROC RANK, we obtain the first 4 columns. We then calculate WOE for each bin and put it in column 5.

Bin Number	Age	Good Accounts	Bad Accounts	WOE
1	10	41	9	-0.061060257
2	20	24	6	0.0689928715
3	30	7	3	0.6079893722
4	40, 48	10	0	Missing
5	Missing	8	3	0.4744579796

Table 4.1: Equal-size Binning

Note that the total number of good accounts g and total number of bad accounts b are 90 and 21, respectively. Hence, $\ln\left(\frac{b}{g}\right) = -1.45529$.

The following SAS data step will perform the WOE transform to transform values of age into WOE.

```
data age2;
      set age1;
      if age = . then M_age = 0.4744579796;
      if age = 10 then M_age = -0.061060257;
      else if age = 20 then M_age = 0.0689928715;
      else if age = 30 then M_age = 0.6079893722;
      else if age > 30 then M_age = . ;
run;
```

Now, let's run logistic regression in SAS as follows.

```
proc logistic data=age3 descending;
      model y = M_age;
run;
```

The output as in Table 4.2 demonstrates that it follows Theorem 3.1.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.4553	0.2794	27.1382	<.0001
M_age	1	1	0.9874	1.0258	0.3112

Table 4.2: Maximum Likelihood Estimate Output 1

Next, let's merge the bin with missing values to bin 3 and run logistic regression again to obtain results Table 4.3.

```
data age4;
   set age2;
   if age > 30 then M_age = 0.6079893722;
run;

proc logistic data=age4 descending;
   model y = M_age;
run;
```

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-1.4565	0.2776	27.5201	<.0001
M_age	1	0.00829	0.9150	0.0001	0.9928

Table 4.3: Maximum Likelihood Estimate Output 2

Obviously, it does not follow Theorem 3.1. Therefore, we have caught binning errors. Indeed, when bin 4 is merged into bin 3, the number of good accounts and number of bad accounts in the new bin should be adjusted to 10 + 7 = 17 and 0 + 3 = 3, respectively. Hence, WOE in the new bin is

$$\ln\left(\frac{\frac{3}{21}}{\frac{17}{90}}\right) = -0.279313823.$$

Let's change WOE and run one more time logistic regression.

```
data age5;
      set age2;
      if age >= 30 then M_age =  −0.279313823;
run;
proc logistic data=age5 descending;
      model y = M_age;
run;
```

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-1.4552	0.2436	35.6961	<.0001
M_age	1	1.0005	1.1640	0.7388	0.3900

Table 4.4: Maximum Likelihood Estimate Output 3

As shown in Table 4.4, this time it follows Theorem 3.1. Note that the slope is not exactly 1 and the intercept is not exactly -1.4553 due to inevitable computational errors. Hence, we have caught and corrected the binning error.

5 Conclusions

In this paper, we have derived an explicit solution to a logistic regression model with one independent variable that has undergone a WOE transformation. We used this explicit solution as a necessary condition for a good binning algorithm and hence provided a simple way to catch binning errors.

References

[1] Albert A. and Anderson, A. (1984). On the existence of maximum likelihood estimates in logistic regression. *Biometrika 71*, 1-10.

[2] Amemiya, T. (1984). Advanced Economics, Cambridge, MA, Harvard University Press.

[3] Fayyad, U. M. and Irani, K. B. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. In: Proceedings of the 13th Int. Joint Conference on Artificial Intelligence, 1022-1027.

[4] Good, I. J. (1985). Weight of Evidence: A Brief Survey. *BAYESWSTATISTICS 2*, 249-270.

[5] Hosmer D. W. and Lemeshow, S. (2000). Applied Logistic Regression. 2nd Ed., John Wiley and Sons, New York.

[6] Refaat, M. (2001). Credit Risk Scorecards: Development and Implementation Using SAS. LULU.COM –USA.

[7] Siddiqi, N. (2006). Credit Risk Scorecards – Developing and Implementing Intelligent Credit Scoring, John Wiley and Sons, New Jersey.

[8] Zeng, G. (2013). Metric divergence measures and information values in credit scoring. *Journal of Mathematics*. Article ID 848271, 10 pages.

[9] Zeng, G. (2014). A Rule of thumb for reject inference in credit scoring. *Mathematical Finance Letters*, Article ID 2.

Received: April 26, 2014