Available online at http://scik.org
Math. Finance Lett. 2014, 2014:2
ISSN 2051-2929

A RULE OF THUMB FOR REJECT INFERENCE IN CREDIT SCORING

GUOPING ZENG AND QI ZHAO

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

Abstract: Credit scoring models are usually developed using the accepted Known Good-Bad applicants, called KGB model. Yet, the KGB model does not represent the entire Through-The-Door population. Reject inference attempts to correct this inherent flaw by using information of the rejected accounts. Augmentation methods are widely used methods of reject inference, among which Fuzzy Augmentation is the most accurate one. In this paper, we first establish an important property of Fuzzy Augmentation: If Fuzzy Augmentation is not incorporated with variable re-selection, it will produce the same results as the KGB model. We then propose a rule of thumb for Augmentation methods. Based on this rule of thumb, we present a two-phase Augmentation. This two-phase method works not only for Machine Learning in Python but also for the traditional approach using SAS. Moreover, it is user friendly in that the user can specify a factor to increase the bad rate of rejected accounts.

Keywords: reject inference; fuzzy augmentation; logistic regression; Machine Learning; Rule of Thumb

2010 AMS Subject Classification: 62-07; 62J12; 97M30; 91G40.

1. Introduction

During the development of a credit scoring model, typically we have only data of accepted (and booked) applicants (or exchangeably accounts) for credit in the past. After operated and observed for some time, these data are used to define the dependent variable y by deriving a

*Corresponding author

Received January 2, 2014

Good or Bad status for each accepted account. On the other hand, the behavior of rejected applicants, if they had been accepted, is unknown and hence ignored. However, when the model using the accepted Known Good-Bad applicants, called KGB model, is applied to all applicants, sample bias, often referred to as reject bias [1] will be gained.

Reject inference [2] attempts to address this kind of reject bias by estimating how rejected applicants would have performed had they been accepted. By using information contained in rejected accounts, reject inference will improve the quality of the scoring model.

The core task in reject inference is to estimate the Good or Bad status, that is, the values of dependent variable y. After the Good or Bad status of each rejected account is known, a new logistic scoring model will be developed using the whole population including both accepted accounts and rejected accounts.

A number of reject inference approaches have been developed [3] over the years. According to [4], reject inference approaches can be classified into 2 categories:

(i) **Simple Assignment** methods: Rejected records are assigned Good or Bad status without using the KGB model.
(ii) **Augmentation** methods: Rejected records are extrapolated with Good or Bad status by extending the KGB model.

Augmentation methods [5] are generally believed better than Assignment methods in that they optimally combine the information of accepted accounts with that of rejected accounts.

In this paper, we shall concentrate on Augmentation methods. We first prove that if the new scoring model with Fuzzy Augmentation does not reselect variables but reuses all the variables in the KGB model, then it will produce the same results as the KGB model. We then propose a rule of thumb for Augmentation methods. Based on this rule of thumb, we present a two-phase Augmentation. This two-phase method works not only for Machine Learning in Python but also for the traditional approach using SAS.

The rest of the paper is organized as follows. In Section 2, we review logistic regression. In Section 3, we introduce Fuzzy Augmentation and explore its true meaning. In Section 4, we propose a Rule of Thumb in reject inference. In Section 5, a novel two-phase Augmentation method is presented. The paper is concluded in Section 6.

2. Basic of logistic regression

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.

To adopt standard notation in logistic regression [6], we use the quantity $\pi(x) = P(y = 1|x)$ to represent the conditional probability that y is equal to 1 given x. It follows that $1 - \pi(x)$ gives the conditional probability that y is equal to zero 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 the product form

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

where β is the vector $(\beta_0, \beta_1, \ldots \beta_p)$. Note that if y_i is known, either 0 or 1, then the 2 terms in the product of (2.3) reduces to only one term as the other term will have a value of 1.

The principle of maximum likelihood states that the solution to the 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

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

The value of β given by the solution to (2.4) is called the maximum likelihood estimate and will be denoted by $\widehat{\beta}$. The maximum likelihood estimate of $\pi(x_{i1}, x_{i2}, \ldots, x_{ip})$ will be denoted by $\hat{\pi}(x_{i1}, x_{i2}, \ldots, x_{ip})$ or simply $\hat{\pi}(x_i)$. It follows from (2.1) that $0 < \hat{\pi}(x_i) < 1$.

The maximum likelihood estimate to (2.4) may not exist, say, in case of complete separation or quasi-complete separation [7]. On the other hand, there exists at most one maximum likelihood estimate [8]. Therefore, if there is a maximum likelihood estimate, it must be unique.

For the weighted logistic regression, let's assume the i-th observation has a positive weight w_i. The weighted likelihood function for logistic regression (2.3) becomes

$$l(\beta) = \prod_{i=1}^{N} \pi(x_{i1}, x_{i2}, \ldots, x_{ip})^{w_i y_i} [1 - \pi(x_{i1}, x_{i2}, \ldots, x_{ip})]^{w_i(1-y_i)}. \tag{2.5}$$

Accordingly, the weighted log likelihood is

$$L(\beta) = \ln[l(\beta)] = \sum_{i=1}^{N}\{w_i y_i \ln[\pi(x_{i1}, x_{i2}, \ldots, x_{ip})] + w_i(1 - y_i)\ln[1 - \pi(x_{i1}, x_{i2}, \ldots, x_{ip})]\}. \tag{2.6}$$

While weights could be any positive numbers, an integral weight implies the frequency of the observation without weights. If an observation is not assigned a weight, it will have a weight of 1 by default. In this sense, (2.4) is a special case to (2.6) with all weights equal 1.

3. Fuzzy Augmentation

Fuzzy Augmentation is a widely used Augmentation method in credit scoring [4]. This method does not simply assign Good or Bad. Instead, it creates weighted Good or Bad using the Good and Bad probabilities calculated from the KGB model. Each record in the rejected accounts is replaced by 2 new records: one with Bad status and the probability of Bad as its weight, the other with Good status and the probability of Good as its weight. The new records and their associated weights, combined with the accepted accounts, are used to develop a new logistic scoring model.

The following theorem illustrates the true meaning of Fuzzy Augmentation as the extension of the KGB model. Before we go through the theorem, let's first state a lemma.

Lemma 3.1. *If* $0 < a < 1$, *then function* $f(x) = a\ln x + (1-a)\ln(1-x)$ *reaches its maximum value in (0, 1) at* $x = a$. *Moreover,* $x = a$ *is the only maximum point of* $f(x)$ *in* $(0,1)$.

Proof. Clearly, the derivatives $f(x)$ in $(0,1)$ satisfies the following inequalities

$$f'(x) = \frac{a}{x} - \frac{1-a}{1-x} = \frac{a-x}{x(1-x)} = \begin{cases} > 0, & if\ x < a \\ 0, & if\ x = a \\ < 0, & if\ x > a \end{cases}$$

Hence, $f(x)$ is strictly increasing in $(0,a)$ and strictly decreasing in $(a,1)$. Since $f(x) \to -\infty$ when $x \to 0$ or 1, $f(x)$ reaches its maximum value in (0, 1) at $x = a$. Moreover, $x = a$ is the only maximum point of $f(x)$ in $(0,1)$. Q.E.D.

Theorem 3.2. If the new logistic scoring model with Fuzzy Augmentation does not reselect variables but reuses all the variables in the KGB model, then it will produce the same maximum likelihood estimate as the KGB model.

Proof. Assume that there are n observations from booked accounts and m observations from rejected accounts with unknown y values. Let $x = (x_1, x_2, \dots, x_p)$ be the vector of p independent variables selected by the KGB model for the n booked accounts. Let $\hat{\beta}$ be the solution to the KGB model and $\hat{\pi}(x_i)$ the maximum likelihood estimate of $\pi(x_i)$.

Therefore, the weighted log likelihood (2.6) for the new logistic model with Fuzzy Augmentation becomes

$$L(\beta) = \sum_{i=1}^{n} \{y_i \ln[\pi(x_{i1}, x_{i2}, \dots, x_{ip})] + (1 - y_i) \ln[1 - \pi(x_{i1}, x_{i2}, \dots, x_{ip})]\} + \sum_{i=n+1}^{n+m} \{\hat{\pi}(x_i) \ln[\pi(x_{i1}, x_{i2}, \dots, x_{ip})]\} + \sum_{i=n+1}^{n+m} \{(1 - \hat{\pi}(x_i)) \ln[1 - \pi(x_{i1}, x_{i2}, \dots, x_{ip})]\} \quad (3.1)$$

Combining the second and third summation in (3.1) for the m rejected accounts, we obtain

$$L(\beta) = \sum_{i=1}^{n} \{y_i \ln[\pi(x_{i1}, x_{i2}, \dots, x_{ip})] + (1 - y_i) \ln[1 - \pi(x_{i1}, x_{i2}, \dots, x_{ip})]\} + \sum_{i=n+1}^{n+m} \{\hat{\pi}(x_i) \ln[\pi(x_{i1}, x_{i2}, \dots, x_{ip})] + (1 - \hat{\pi}(x_i)) \ln[1 - \pi(x_{i1}, x_{i2}, \dots, x_{ip})]\}. \quad (3.2)$$

Then, $\hat{\beta}$ is the solution to（3.2）without the second part for the rejected accounts. Applying Lemma 3.1. to each of the m items in the second summation of（3.2), we see that the m

items will each reach its maximum when $\pi(x_{i1}, x_{i2}, \dots, x_{ip}) = \hat{\pi}(x_i)$. Since

$\pi(x_{i1}, x_{i2}, \dots, x_{ip}) = \hat{\pi}(x_i)$ when $\beta = \hat{\beta}$, it follows that $\beta = \hat{\beta}$ maximizes each item and

hence maximizes the second summation. Since $\beta = \hat{\beta}$ already maximizes the first summation, $\beta = \hat{\beta}$ maximizes $L(\beta)$. Since (3.2) has at most one maximum likelihood estimate, $\beta = \hat{\beta}$ is the only solution to (3.2). Q.E.D.

To implement Fuzzy Augmentation, one needs to reselect variables for the new logistic scoring model for the whole population after applying the KGB model. This can be done by means of Information Value or Metric Divergence measures L_∞, L_2 and L_1 [9]. Due to reject inference, some variables in the KGB model may be out and some new variables may be in. Since the new model for the whole population and the KGB model are likely to select different variables, they are likely to have different maximum likelihood estimates.

4. A Rule of Thumb for reject inference

Suppose we need to develop a new Augmentation method. Let's assume that there are n observations from booked accounts and m observations from rejected accounts with unknown y_i. Let $\hat{\beta}$ be the solution to the KGB model and $\hat{\pi}(x_i)$ the maximum likelihood estimate of $\pi(x_i)$.

Separating booked accounts and rejected accounts from Equation (2.4) with $N = n + m$ yields 2 summations

$$L(\beta) = \sum_{i=1}^{n}\{y_i \ln[\pi(x_{i1}, x_{i2}, \dots, x_{ip})] + (1 - y_i)\ln[1 - \pi(x_{i1}, x_{i2}, \dots, x_{ip})]\} +$$
$$\sum_{i=n+1}^{n+m}\{y_i \ln[\pi(x_{i1}, x_{i2}, \dots, x_{ip})] + (1 - y_i)\ln[1 - \pi(x_{i1}, x_{i2}, \dots, x_{ip})]\}. \quad (4.1)$$

Replacing y_i for rejected accounts with their respective maximum likelihood estimate $\hat{\pi}(x_i)$, we obtain

$$L(\beta) = \sum_{i=1}^{n}\{y_i \ln[\pi(x_{i1}, x_{i2}, \ldots, x_{ip})] + (1-y_i)\ln[1-\pi(x_{i1}, x_{i2}, \ldots, x_{ip})]\} + \sum_{i=n+1}^{n+m}\{\hat{\pi}(x_i)\ln[\pi(x_{i1}, x_{i2}, \ldots, x_{ip})] + (1-\hat{\pi}(x_i)\ln[1-\pi(x_{i1}, x_{i2}, \ldots, x_{ip})]\}. \quad (4.2)$$

Then (4.2) is the same as (3.2). From the proof of Theorem 3.2., one may jump to conclusion that this new model will yields the same results as the KGB model. Indeed, this is not true simply because $\hat{\pi}(x_i)$ is not dichotomous but a real number between 0 and 1. However, (4.2) can be treated as the weighted log likelihood.

On the other hand, we note that the expected values of y_i's are $\hat{\pi}(x_i)$ for all rejected accounts in Fuzzy Augmentation. This motivates us to think the following Rule of Thumb when developing a new Augmentation method.

A Rule of Thumb in Reject Inference: When developing an Augmentation method for reject inference, it is better to make the expected values of the estimated y_i's for rejected accounts equal their respective predicted probabilities from the KGB model.

5. A two-phase augmentation method for Machine Learning in Python

It is easy to implement Fuzzy Augmentation in the traditional approach using SAS simply because Proc Logistic in SAS has a Weight option. Yet, to the best knowledge of the authors, there are no existing packages for weighted logistic for Machine Learning in Python.

5.1. Two-Phase Augmentation

Let b denote the bad rate of booked accounts, that is,

$$b = \frac{\sum_{i=0}^{n} y_i}{n}.$$

Phase I: Basic Model

Step 1: Use the KGB model to find the bad probabilities of rejected accounts. Let p_i be the probability of bad for a rejected account i.

Step 2: Generate a random number r_i between 0 and 1 for each rejected account i. Assign a Good or Bad status to each rejected account as follows:

$$y_i = \begin{cases} 1, & \text{if } r_i \leq p_i \\ 0, & \text{if } r_i > p_i. \end{cases}$$

Step 3: Calculate the bad rate of the rejected accounts as follows

$$\bar{b} = \frac{\sum_{i=1}^{m} y_i}{m}.$$

If it is high enough, say, 2 times of the bad rate of the booked accounts, start to develop a new model for all the accounts. Otherwise, continue the following steps to adjust the bad rate of the rejected accounts.

Phase II: Extended Model

Step 4. Regenerate a random number r_i between 0 and 1 for each rejected account i. Assign a Good or Bad status to each rejected account as follows:

$$y_i = \begin{cases} 1, & \text{if } r_i \leq \dfrac{\alpha b p_i}{\bar{b}} \\ 0, & \text{otherwise,} \end{cases}$$

where $\alpha > 1$ and $\alpha b < 1$.

Step 5. Start to develop a new model for all the accounts.

5.2. Analysis of two-phase augmentation

We shall now analyze the correctness of the two-phase Augmentation method.

Theorem 5.1. *The basic model in Phase I of the two-phase Augmentation method follows the rule of thumb in reject inference.*

Proof. As is known, a uniformly distributed random variable X on a probability space $\{[0,1],\mathcal{F},P\}$ is a real-valued function $X:[0,1]\to\mathcal{R}$ by $X(\omega)=\omega$. Here, $\mathcal{F}$ is Borel sigma field which includes all the subsets of [0, 1] generated by all subintervals $[a,b]$ of $[0,1]$ such that $0\le a\le b\le 1$ after finite set operations (union, intersections, complements and differences). It can be proved that any set in $\mathcal{F}$ is a finite union of disjoint intervals (closed, open or half-closed). P is a probability measure defined as

(i) $P([a,b])=b-a.$

(ii) $P(A)=\sum_{i=1}^{k}(b_i-a_i),$ for any $A=\bigcup_{i=1}^{k}[a_i,b_i]\epsilon\mathcal{F}.$

Since for each i, random number r_i follows a uniform distribution X in [0, 1], we have

$$P(X\le p_i)=p_i,\qquad P(X\ge p_i)=1-p_i.$$

Next, for each i, y_i is a discrete random variable on the same probability space $\{[0,1],\mathcal{F},P\}$ as the uniform random variable defined by: $y_i:[0,1]\to\mathcal{R}$

$$y_i=\begin{cases}1, & \text{if } X(\omega)=\omega\le p_i\\ 0, & \text{if } X(\omega)=\omega>p_i.\end{cases}$$

Therefore, the expected value of $y_i=1\times p_i+0\times(1-p_i)=p_i.$ Hence, the basic model follows the Rule of Thumb in reject inference. Q.E.D.

Theorem 5.2. *For the extended model in Phase II of the two-phase Augmentation method, the average bad rate of the rejected accounts is* α *times of the bad rate of the booked accounts.*

Proof. We shall adopt the same notation about uniformly distributed random variable X. Since for each $i = n+1, n+2, \ldots, n+m$, random number r_i is a uniformly distributed random variable X in [0, 1], we have

$$P(X \le r_i) = \frac{abp_i}{\bar{b}}, \qquad P(X \ge r_i) = 1 - \frac{abp_i}{\bar{b}}.$$

For each $i = n+1, n+2, \ldots, n+m$, y_i is a random variable on $\{[0,1], \mathcal{F}, P\}$ defined by:

$y_i: [0,1] \to \mathcal{R}$

$$y_i = \begin{cases} 1, & \text{if } X(\omega) = \omega \le \dfrac{abp_i}{\bar{b}} \\ 0, & \text{if } X(\omega) = \omega > \dfrac{abp_i}{\bar{b}}. \end{cases}$$

Then the expected value $E(y_i)$ of y_i can be found as

$$E(y_i) = 1 \times \frac{abp_i}{\bar{b}} + 0 \times \left(1 - \frac{abp_i}{\bar{b}}\right) = \frac{abp_i}{\bar{b}}.$$

Now define the actual bad rate of the reject accounts by Z, then

$$Z = \frac{\sum_{i=n+1}^{n+m} y_i}{m}.$$

Z is also a random variable on $\{[0,1], \mathcal{F}, P\}$. Since $y_i, i = n+1, n+2, \ldots, n+m$ are unknown until they are assigned, we turn to its expected value $E(Z)$, that is, the average bad rate of the reject accounts.

$$E(Z) = \frac{\sum_{i=n+1}^{n+m} E(y_i)}{m} = \frac{\sum_{i=n+1}^{n+m} \frac{abp_i}{\bar{b}}}{m} = \frac{ab}{\bar{b}} \frac{\sum_{i=n+1}^{n+m} p_i}{m} = \frac{ab}{\bar{b}} \bar{b} = ab.$$

Hence, the average bad rate of the rejected accounts is a times of the bad rage of the booked accounts. Q.E.D.

Remark 5.3. It follows from Theorem 5.2. that if $\alpha = \frac{\bar{b}}{b}$, then the bad rate of rejected accounts is $\bar{b}$. In this case, Phase II will have the same results as Phase I. Therefore, we may set the default value of α to $\frac{\bar{b}}{b}$.

Remark 5.4. Usually, the bad rate of rejected accounts after Phase I is much higher than that of booked accounts. In this case, we don't need Phase II. In case Phase II is needed, we may adjust the bad rate of rejected accounts to meet the need.

Remark 5.5. This two-phase Augmentation method has several advantages over other Augmentation methods. It works not only for Machine Learning in Python but also for the traditional approach using SAS. It is user friendly in that the user can specify a factor α to increase the bad rate of rejected accounts.

6. Conclusions

In this paper, we first proved an important property of Fuzzy Augmentation: If Fuzzy Augmentation is not incorporated with variable re-selection, it will produce the same results as the KGB model. We then proposed a Rule of Thumb in reject inference. Based on this Rule of Thumb, we presented a novel two-phase Augmentation method. This two-phase augmentation method works not only for Machine Learning in Python but also for the traditional approach using SAS. Moreover, it is user friendly in that the user can specify a factor to increase the bad rate of rejected accounts.

Conflict of Interests

The authors declare that there is no conflict of interests.

REFERENCES

[1] J. L. Banasik, J. N. Crook, J.N. and L. C. Thomas, Sample selection bias in credit scoring models, Journal of the Operational Research Society. 54 (2003), 822–832.

[2] D. J. Hand and W. E. Henley, Can reject inference ever work? IMA Journal of Mathematics Applied in Business and Industry. 5 (1993), 45–55.

[3] J. N. Crook and J. L. Banasik, Does reject inference really improve the performance of application scoring models? Journal of Banking & Finance. 28 (2004), 847-874.

[4] M. Refaat, Credit Risk Scorecards: Development and Implementation Using SAS, LULU.COM - USA, 2011.

[5] J. Banasik and J. Crook, Reject inference, augmentation, and sample selection, European Journal of Operational Research. 183 (2007), 1582–1594.

[6] D. W. Hosmer and S. Lemeshow, Applied Logistic Regression, 2nd edition, John Wiley & Sons, Inc., 2000.

[7] A. Albert and J. A. Anderson, On the Existence of Maximum Likelihood Estimates in Logistic Regression, Biometrika. 71(1984), 1-10.

[8] T. Amemiya, Advanced Econometrics. Cambridge, MA: Harvard University Press, 1985.

[9] G. Zeng, Metric Divergence Measures and Information Values in Credit Scoring, Journal of Mathematics. 2013 (2013), Article ID 848271, 10 pages.