IOSR Journal of Economics and Finance (IOSR-JEF)
e-ISSN: 2321-5933, p-ISSN: 2321-5925.Volume 3, Issue 3. (Mar-Apr. 2014), PP 64-68
www.iosrjournals.org

Model Validation of a Credit Scorecard Using Bootstrap Method

Dilsha M[1], Kiruthika[2]

[1]*(Department of Statistics, Pondicherry University, India)*
[2]*(Department of Statistics, Pondicherry University, India)*

Abstract: *The validation of a credit risk scorecards is very important. Lenders therefore need a proper validation methodology to convince their supervisors that their credit scoring models are performing well. In this paper we propose a simple validation methodology by using bootstrap. We compare this methodology by applying it to a default model of microfinance credit scorecard.*

Keywords: *Bootstrap, Microfinance, credit scorecard*

I. Introduction

The main principle of a credit scoring system is assigning to each borrower a score in order to separate the bad borrowers (defaulters) and the good borrowers (non-defaulters). For example, a borrower with high estimated default probability is given a high score. So, a scoring system can be seen as a classification tool in the sense of providing indications of the borrower's possible status in the future. This procedure is commonly called discrimination. Thus, the discriminatory power of a scoring system denotes the model's ability to discriminate defaulters from non-defaulters. Assessing the discriminatory power is one of the major tasks in validating a credit scoring model.

These days, a lot of emphasis has been given to the validation of the internal rating system. Here the validation includes both the assessment of model discriminatory power and calibration. While the discriminatory power of a scoring model depends on the difference of the score distribution of the defaulters and non-defaulters, the calibration of a scoring system depends on the difference of the estimated probability of default and the observed default rates. So for a proper modelling of a credit scorecard requires a large number of sample data. If the sample size ‘n’ is small then validating the data is very difficult. In this case, we found that the model suffer either substantial bias or variability. Validation is obviously not only a statistical exercise. Managerial judgment and a qualitative analysis of the model are also highly important. However, the initial validation will primarily be technical and model based. Moreover, statistical validation is needed to obtain scientific rigor and a common yardstick for the validation exercise. For these reasons, this article will focus on a quantitative validation technique and propose a statistical validation methodology. The main focus of this paper is how to adjust the parameter estimate of the model after running the bootstrap method and then checking the robustness of the adjusted parameter.

This paper divided in to four sections. First an introduction to the most general concepts of the credit scoring methodologies and followed by the statistical techniques like variable selection, logistic regression and statistical model validation is described. The main objective of this paper is to find out a validation methodology using bootstrap which is described in Section 3.3. The methodology of validating a logit model applied in medical science is used in this paper which is based on Harrell’s c-index (Harrell 2001). In the last section, some descriptive statistics of the datasets used in the analysis and the experimental results are shown. Finally, the conclusion is presented.

II. Methodologies For Credit Scoring

The three main approaches for credit scoring is judgmental, statistical and non-statistical (Thomas, 2000). This paper focuses on the statistical approach, which is based on historical data and includes methodologies as discriminant analysis and logistic regression.

Discriminant analysis is a computationally efficient procedure, but is hampered by the assumption of normality distributed data (Sharma 1996; Vogelgesgang, 2003). As the models presented in this paper include multiple dummy variables, the normality assumption is violated and therefore discriminant analysis has not been adopted.

Logistic regression employs maximum likelihood estimators which require computationally more demanding procedures than discriminant analysis and linear regression do. However, logistic regression models are not constrained by the assumption of normally distributed data (Sharma, 1996) and they model a probability; their output is a percentage term which is directly interpretable and usable to perform operational actions such as setting cut-off values; Due to these benefits, logit model have been adopted in this paper.

III. Explanatory Variable Treatment And Selection

Once data has been prepared and a statistical model has been chosen, the next step is to decide on the treatment of the explanatory variables. Afterwards, the explanatory variable selection process needs to be considered. Several of the explanatory variables in a credit scoring context are typically categorical (e.g. purpose, previous debt). According to Thomas (2000), there are two options to implement categorical variables in a scoring model. First, a binary (dummy) variable can be created for each possible category of an explanatory variable. Such implementation permits the modeling of non-linear behavior. Crook et al. (1992) note that such a dummy approach can considerably reduce the degrees of freedom available in the model. In addition, near-singularity problems might arise for dummy coded variables when executing the logistic regression. Therefore, another approach, Weight of Evidence (WoE), works with one variable for all categories of an explanatory variable (Crook et al., 1992; Thomas, 2000; Hand and Henley,1997).

The WoE measure the strength of each attributes, or grouped attributes in separating good and bad accounts. It is a measure of difference between the proportion of goods and bad in each attribute (i.e, the odds of a person with that attribute being good or bad). WoE is based on odds calculation and the details are given below.

Let b_i defined as the number of defaulted loans that belong to the i-th group, g_i is defined similarly for the non-defaulted loans. B and G are the total number of defaulted and non-defaulted loans present in the whole sample, defined as:

$$B = \sum_{i=1}^{n} b_i \ and\ G = \sum_{i=1}^{n} g_i$$

In the first WoE step, to avoid over fitting, categories are put together in n groups based on similarity of gi/(gi + bi). Next, each of the newly created n groups receives a coding based on its distribution of defaulted and non-defaulted loans. Hence, every weight of evidence variable is composed of n values, one for each of the groups. Boyle et al. (1992) propose different implementations for the coding procedure:

(i) gi/bi, (ii) gi/(gi + bi), (iii) bi/(gi + bi), (iv) log(gi/(gi + bi)), (v) ln(gi/bi) + ln(B/G).

This study we have used ln(gi/bi) + ln(B/G). A potential weakness of the weight of evidence approach is that the explanatory variable coding is based on the dependent variables, which might cause over fitting on the sample data and result in inferior performance when tested out of-sample.The categorization of a continuous variable can be chosen so that the default risk in the created categories is as homogeneous as possible. Based on similarity of gi/(gi +bi), the individual values of the continuous variable can be grouped (Crook et al.,1992). Grouping values together must be done such that the aggregated values appear sufficiently often in the data set in order to obtain statistically robust results (Boyle et al., 1992). This aggregation process creates groups for each originally continuous variable and is applied for model presented in this paper.

3.1. Regression Model

A regression model is commonly used to study relationship between multiple independent and dependent variables and to determine significant independent variables related to a dependent variable. This model is also able to describe the magnitude and direction of independent variables effect on a dependent variable (Chen & Hughes, 2004). Basically, there are two common categories of regression models: the linear regression model and the logistic regression model. The decision to choose linear regression or logistic regression depends on the measurement scale of a dependent variable. If a dependent variable is expressed on an interval scale, a linear regression is more appropriate to be used. If a dependent variable is a binary/dichotomous data, a logistic regression provides more meaningful results (Agresti, 2002). In this paper, since the dependent variable is binary logistic regression is used for scorecard development.

3.1.1 Logistic Regression

The logistic regression model relates to *y* and *x* can be represented mathematically as follows.

$$\Pr((y = 1)/x) \quad = \frac{1}{1 + e^{-(\beta_0 + \Sigma\, \beta_i x_i)}}$$

The unknown regression coefficients β_i, which have to be estimated from data, are directly interpretable as log-odds ratios, or in terms of exp (β_i), as odds ratios. The assumption is the predictor variables are related in a linear manner to the log odds {log[p/1-p]} for the outcome of interest. Variables are often selected for inclusion in logistic regression models using some form of the backward or forward stepwise regression technique. Widely accepted criteria such as the Hosmer-Lemeshow statistic have been developed for assessing the goodness of fit for logistic regression models. This method is widely used to develop credit risk scorecards in order to predict the probability of a customer having a good payment habit if a loan is granted. Although there are other techniques that could increase the predictive power of the models, the logistic regression has two strong features in its favor namely, simplicity on the model developments and ease of interpretability.

Including all variables would make the model unnecessarily large and deter clients when confronted with the required number of questions. Therefore authors typically adopt explanatory variable selection. Hand and Henley (1997) describe three approaches on this matter. First, expert knowledge can be used to select the right variables. Secondly, statistical procedures as the forward and backward selection based on R^2 can be implemented. A combination of the forward and backward approaches, the stepwise approach, also exists. As a third approach, Henley and Hand propose to select variables by using a measure which indicates the difference between the distributions of the defaulted and non-defaulted loans on that variable. Other authors, such as Verstraeten and Van den Poel (2005) also refer to the importance of the Receiver Operating Characteristic (ROC) Curve and its summary index Area under the ROC Curve (AUC) in the explanatory variable selection process. The ROC Curve gives a graphical representation of the discriminatory power of a scoring system. Baesens et al. (2009) propose a heuristic variable selection procedure, based on AUC, which removes in each consecutive step the variable which causes the smallest decrease in AUC. Based on an expert decision, the tradeoff between strong AUC performance and number of variables is established. In this paper, the stepwise selection procedure is used for the variable selection.

3.2 Statistical Model Validation

In the existing literature (Harrell (2001), Basel Committee on Banking Supervision (2005) and Engelmann and Rauhmeier (2006)) models are validated by determining the discrimination and calibration of the model. A model's discrimination is the ability to separate between defaulters and non-defaulters. Calibration is the ability of the model to make unbiased estimates of the outcome. We say that a model is well calibrated when a fraction of *p* of the events we predict, with a probability *p* actually occur. Discrimination and calibration both compare the estimated probabilities with the observed frequency of default in the data set. So by measuring discrimination and calibration the probability of defaults are validated. However, validation can be more rigorous since the parameters of the model ($\hat{\beta}$) can also be validated. We validate the parameters by means of reproducibility of research, stability of parameters and choice of functional form. Besides we describe bootstrap to validate the probability of defaults as well as the parameters.

3.2.2 Stability of Parameters

There are two types of stability, stability over time and stability over groups. Often models are intended to be used for predictions, but predictions are only valid if parameters are stable over time. In general we are often interested in stability over time for a sub vector of the parameter vector β_i. A stable model requires well determined coefficients with high confidence and similar results on performance characteristics if tested in and out of sample.

In this paper we have used bootstrap approximation method for validating the data. The bootstrap approximation is performed using the following steps.

1. The test statistic LR_{max} is calculated, where LR_{max} be the likelihood ratio test with the highest value.
2. N bootstrap samples are generated from the original data and the model parameters are estimated under H_0.
3. For each bootstrap sample the test statistic LR_{max} is calculated, this results in the so-called bootstrap distribution.
4. The p-value is approximated by the fraction of bootstrap LR_{max} values larger than the LR_{max} obtained using the observed data.

Diebold and Chen (1996) found that the second approximation using the bootstrap distribution outperforms the asymptotic distribution; therefore we use the bootstrap approximation in the application.

3.3 Out-of-sample Performance and Bootstrap

The statistics defined above can be applied to the development set to determine the performance of the model. However, we want to determine the performance of the model for future predictions. Using the same data both to develop the model and to determine the performance of the model will result in an overestimation of the performance for future predictions. If the performance is determined on the development set the estimated performance will be too optimistic. To correct for this optimism out-of-sample performance and bootstrap methods can be applied. So, we are interested in how well the model performs on a different set than the development set. Hence we need two data sets to determine the out-of-sample performance, a development sample and a test sample. First, the model is developed based on the development sample. Second, the test sample is used to determine the out-of-sample performance of the model by means of calculating the discrimination and calibration of the model.

In general we can split the original data into a development and a test sample in two ways. This results in two types of out-of-sample performance, that is, out-of-sample performance within the time period and out-of-sample performance outside the time period. To determine the out of-sample performance within the time period a subset of the complete dataset is used in model development and hence the development set contains observations over T periods. The remaining data also contains observations over T periods and is used to determine the out-of-sample performance of the model. Out-of-sample performance outside the time period means that the data is split in the following way. The observations in the first T –q periods are used to develop the model and the observations in the last q periods are used to determine the out-of-sample performance.

The disadvantage of out-of-sample performance is that the size of the sample used to develop the model is smaller than the original sample of size. The bootstrap method overcomes this problem. The bootstrap method first generates N bootstrap samples. A bootstrap sample is a sample with replacement of size N drawn from the original sample. On each of these bootstrap samples the model is estimated. The N fitted models are applied to the original sample to give N values of a discrimination or calibration measure. The overall accuracy is the average of the N measures. This simple bootstrap method turns out not to work very well. Efron and Tibshirani (1993) describe an enhanced method that works better than the simple method. It is shown that this enhanced method performs better than the simple method (see for example Gong (1986) or Efron (1990)). First N bootstrap samples are drawn and N models are estimated using the bootstrap samples. The fitted models are applied to the original sample to give N measures. The fitted models are also applied to the bootstrap samples (used to fit the model) to give N measures based on the bootstrap samples used to fit the model. The so-called optimism is calculated for each bootstrap sample by taking the difference between the measure based on the original sample and the measure based on the bootstrap sample. This results in N values of the optimism. The overall optimism is the average of the N values of optimism. To determine the discrimination or calibration of the final model, the overall optimism is subtracted from the measure calculated on the final model which is fitted based on the original sample.

IV. Data And Model Description

The data for our study is collected from various self-help groups in Vazhayour and Ramanatukara panchayath in Kerala, India. A self-help group is village based financial intermediary usually composed of ten to twenty local women. Members make small regular savings contribution over a few months until there is enough capital in the group to begin lending. Funds may then be lent back to the members or to others in the village for any purpose. Our dataset contains customer information such as personal characteristics, disposable income, person’s occupation, length of employment, home ownership, characteristics of the current financial operation, variables related to the macroeconomic context, and any delays in the payment of a microcredit fee. In this scenario, if a customer is not paying continuously three weeks then the account is defined as bad, all other accounts are considered as good. To restrict the number of variables we used information value cut of 0.1 and restricted 15 variables for the analysis.

The weights of evidence transformed variables are used for the model development. For variable selection stepwise logistic regression was used initially. Table 1 shows the development model. Four variables (Job, Repaid the loan, Age and income) are significant.

Table 1: Logistic model Parameter estimate

Parameter	DF	Estimate	Standard Error	Wald Chi-square	Pr > ChiSq
Intercept	1	2.0686	0.01906	249.7266	<.0001
JOB	1	1.3432	0.05142	46.8232	0.009
REPAID_THE_LOAN	1	0.8375	0.0348	135.7923	<.0001
AGE	1	1.218	0.02815	196.3979	<.0001
INCOME	1	3.8146	0.0487	114.9394	<.0001

Stability of the variable is measured by p-values of coefficients and bootstrap method. Table 2 shows how many times the variable is significant in bootstrap sampling out of 1000.

Table 2: Bootstrap method number of time significant

Variable	Number of times significant
JOB	959
REPAID_THE_LOAN	974
AGE	982
INCOME	965

For measuring the accuracy we have used concordance. The model concordance is 81.89. Since the weight of evidence procedure is used to develop the model there is a chance to over predict the concordance in bootstrap sampling. So we are finding out the unbiased estimate of concordance. Fig.1 shows the histogram of the concordance measures of 1000 bootstrap sampling. The range of the concordance is 5.5. The unbiased concordance is 81.8. It is very close to model concordance. So it proves we can use bootstrap method for accuracy. Similarly one can find the unbiased parameter estimate. For estimating unbiased parameter, we used 1000 bootstrap estimate of the variable. For example, consider the variable income. We can draw a histogram of the estimate of income using 1000 bootstrap and one can find out the unbiased estimate the same way which is described above for concordance measures. The unbiased estimate of the variables is shown in Table 3.

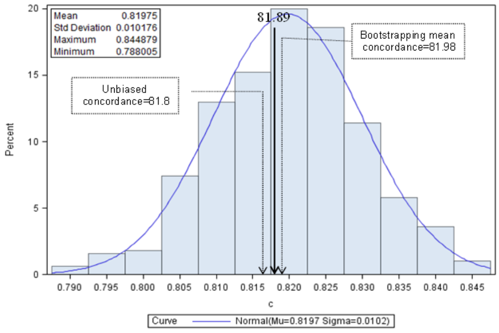

Figure 1: Distribution of C-index

Table 3: Unbiased Parameter Estimate

Variable	Lower limit	Upper limit	Unbiased estimate
JOB	1.098	1.432	1.3314
REPAID_THE_LOAN	0.8525	0.8502	0.8379
AGE	1.0821	1.2264	1.2059
INCOME	3.7936	3.8587	3.8013

V. Conclusion

In this paper we have shown that the bootstrap method can be used for validating a credit scorecard. The proposed methodology mainly can be used when the sample size is small. Usually the out of sample size data is very small compared to the development data. One can use this method in the out of time data in the scorecard development process. Also we have given a basic idea and how to develop more accurate scorecard in this paper. Although this procedure has only been tested in a credit scorecard model selection context, this simple and time saving method could easily be extended to other contexts of nonlinear regression, classification, etc., where computation time and complexity play a role.

References

[1] Frank. E. Harrell, *Regression Modeling Strategies*(New York: Springer, 2001)

[2] Thomas. L, A Survey of Credit and Behavioural Scoring; Forecasting financialrisk of lending to consumers, *International Journal of Forecasting*, 16(2), 2000, 149 -172.

[3] Sharma. S, *Applied Multivariate Techniques* (New York, NY: John Wiley Sons, 1996).

[4] Vogelgesang, U, Microfinance in Times of Crisis: The Effects of Competition, Rising Indebtness, and Economic Crisis on Repayment Behaviour, World *Development*, 31(12), 2003, 2085-2114.

[5] Crook, J., Hamilton, R. and Thomas, L. A comparison of discriminations underalternative definitions for credit default. in L. Thomas, J. Crook and D. Edelman (eds),Credit Scoring and CreditControl. *Oxford: Oxford University Press*, 1992,217-245.

[6] Hand, D. and Henley, W, Statistical Classification Methods in Consumer CreditScoring*: A Review.Journal of the Royal Statistical Society Series A* (Statistics in Society), 160(3), 1997, 523-541.

[7] Boyle, M., Crook, J., Hamilton, R. and Thomas, L, Methods for credit scoring applied to slow payers. in L. Thomas, J. Crook and D. Edelman (eds), Credit Scoringand Credit Control. *Oxford: Oxford University Press*, 1992, 75-90.

[8] Chen, C.-K, & Hughes, J, Using ordinal regression model to analyse student satisfaction questionnaire, *IR Applications*, 1, 2004, 1-13.

[9] Agresti, A, *Categorical data analysis* (New Jersey: Wiley-Interscience. Second ed.,2002).

[10] Verstraeten, G. and Van den Poel, D, The impact of sample bias on consumercredit scoring performance and profitability, *Journal of the Operational Research Society*, 56(8), 2005, 981-992.

[11] Baesens, B., Van Gestel, T. and Thomas, L*Credit Risk Management: BasicConcepts*.(Oxford: Oxford University Press, 2009).

[12] Engelmann, B. and R. Rauhmeier *The Basel II Risk Parameters:Estimation, Validation, and Stress Testing*(Heidelberg: Springer, 2006).

[13] Diebold, Francis X. and Celia Chen, Testing structural stability with endogenous breakpoint: A size comparison of analytic and bootstrap procedures, *Journal of Econometrics*, 70, 1996, 221–241.

[14] B. Efron, R. J. Tibshirani, *An introduction to the bootstrap*(Chapman & Hall., 1993).

[15] Gong, Gail, Cross-validation, the jackknife, and the bootstrap: Excess error estimation in forwardlogistic regression, *Journal of the American Statistical Association*, 81, 1986, 108–113.

[16] Efron, Bradley, More efficient bootstrap computations, Journal *of the American Association* 85, 1990, 79–89