Measuring Scorecard Performance

Zheng Yang, Yue Wang, Yu Bai , and Xin Zhang

Colledge of Economics, Sichuan University
Chengdu, Sichuan, 610064, China
Yangzheng9@126.com

Abstract. In this paper, we look at ways to measure the classification performance of a scoring system and the overall characteristics of a scorecard. We stick to the idea that we will measure the scoring system by how well it classifies, which are still problems in measuring its performance. This is because there are different ways to define the misclassification rate mainly due to the sample that we use to check this rate. If we test how good the system is on the sample of customers we used to build the system, the results will be better than that we did the test on another sample. This idea is illustrated in this paper. Two measures, Mahalanobis distance and KS score, are used in the paper.

Keywords: credit-scoring systems, measuring scorecard, classification, holdout, mahalanobis distance, KS score.

1 Introduction

Having built a credit or behavioral scorecard, the obvious question is, ”How good is it ?” this begs the question of what we mean by good. The obvious answer is in distinguishing the good from the bad because we want to treat these groups in different ways in credit-scoring systems---for example, accepting the former for credit and rejecting the latter. Behavioral scoring systems are used in a more subtle way, but even if we stick to the idea that we will measure the scoring system by how well it classifies, there are still problems in measuring its performance. This is because there are different ways to define the misclassification rate, mainly due to the sample that we use to check this rate. If we test how good the system is on the sample of customers we used to build the system, the results will be must better than if we did the test on another sample. This must follow because built into the classification system are some of the nuances of that data set that do not appear in other data sets. Thus section 2 looks at how to test the classification rate using a sample, called the holdout sample, separate from the one used to build the scoring system. This is a very common thing to do in the credit-scoring industry because of the availability of very large samples of past customers, but it is wasteful of data in that one does not use all the information available to help build the best scoring system. There are times, however, when the amount of data is limited, for example, when one is building a system for a completely new group of customers or products. In the case, one can test

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 900–906, 2004.

the validity of the system using methods that build and test on essentially the same data set but without causing the misclassification errors to be optimistically biased. There are a number of standard ways to describe how different two populations are in their characteristics. These can be used in the case of scorecard-based discrimination methods to see how different the scores are for the two groups of good and bad. These measure how well the score separates the two groups, and in section 3, we look at two such measures of separation-the Mahalanobis distance and Kolmogorov-Smirnov statistics. These give a measure of what is the best separation of the groups that the scorecard can make.

2 Error Rates Using Holdout Samples and 2 × 2 Tables

We defined the decision making in a credit-scoring system as follows: X=(X_1,X_2,...,Xp) are the application variables, and each applicant gives a set of answers **x** to these variables and thereafter is found to be either good(G) or bad(B). Assuming that the application characteristics are continuous (a similar analysis works for the discrete case), let f(x) be the distribution of application characteristics, f(G/x) be the probability of being a good if the application characteristics were **x**, and f(B/x)=1 − f(G/x) be the probability of being a bad if those are the characteristics.

The optimal or Bayes error rate is the minimum possible error rate if one knew these distributions completely over the whole population of possible applicants.

$$e(Opt) = \int \min\{f(B|x), f(G|x)\} f(x)dx \tag{1}$$

Any given credit-scoring system built on a sample S of n consumers estimates the functions f(G/x)and f(B/x)and in light of this defines two regions of answers A_G and A_B ,in which the applicants are classified as good or bad. The actual or true error for such a system is then defined as

$$e_S(Actual) = \int_{A_G} f(B|x)f(x)dx + \int_{A_B} f(G|x)f(x)dx \tag{2}$$

This is the error that the classifier built on the sample S would incur when applied to an infinite test set. The difference occurs because one has used only a finite sample S to build the estimator. One is usually interesting in estimating e_S (Actual), but if one had to decide what system to use before seeing the data and adapting the system to the data, then one would take the expected error rate e_n(Expected),which is the expectation of e_S(Actual) over all samples of size n.

The difficulty in calculating e_S(Actual)is that one does not have an infinite test set on which to calculate it and so one has to use a sample S* to estimate it on. Hence one calculates

$$e_S(S^*) = \int_{A_G} f(B|x)fs^*(x)dx + \int_{A_B} f(G|x)fs^*(x)dx \tag{3}$$

where fs*(x) is the distribution of characteristics **x** in the sample S*.

The obvious thing is to check the error on the sample on which the classifier was built and so calculate e_S(S). Not surprisingly, this underestimates the error considerably, so e_S(S) < e_S(Actual). This is because the classifier has incorporated in it

all the quirks of the sample S even if these are not representative of the rest of the population. Hence it is far better to test the data on a completely independent sample S^*. This is called the holdout sample, and it is case that the expected value of $e_S(S^*)$ over all samples excluding S is e_S(Actual). This means that this procedure gives an unbiased of the actual error.

In credit scoring, the cost of the two errors that make up the error rate are very different. Classifying a good as a bad means a loss of profit L, while classifying a bad as a good means an expected default of D, which is often considerably higher than L. Thus instead of error rates one might look at expected loss rates, where the optimal expected loss l(Opt) satisfies

$$l(Opt) = \int \min \left\{ Df(B|x),\ Lf(G|x) \right\} f(x)\,dx \tag{4}$$

In analogy with (2) and (3), the actual or true loss rate for a classifier based on a sample S is

$$l_S(Actual) = \int_{A_G} Df(B|x) f(x) dx \int_{A_B} Lf(G|x) f(x) dx \tag{5}$$

While the estimated rate using a test sample S^* is

$$l_S(S^*) = \int_{A_G} Df(B|x) f_{S^*}(x) dx \int_{A_B} Lf(G|x) f_{S^*}(x) dx \tag{6}$$

Bayes's theorem will confirm that the expression for the actual loss rate in (5).

So how does one calculate $e_S(S^*)$ and $ls(S^*)$, having built a classifier on a sample S and having a completely independent holdout sample S^* available? What we do is to compare the actual class G or B of each customer in the sample S^* with the class that the scorecard predicts. The results are presented in a 2×2 table called the confusion matrix (see Table 1) which gives the numbers in each group.

For example, we might have the confusion matrix given in Table 2. In this sample, n, n_G, and n_B (1000,750,250) are fixed as they describe the sample chosen. Thus really only two of the four entries g_G, b_G, g_B, and b_B are independent. The actual error rate is calculated as $(b_G+g_B)/n$=250/1000=0.25. If the losses are L=100, D=500, then their actual loss per customer is $(Lb_G+Dg_B)/n$=(100*150)+(500*100)/1000=65. One can use confusion matrices to compare systems or even to try to decide on the best cutoff score when a scorecard had been developed. In the latter case, changing the scorecard in the example in Table 2 may lead to the confusion matrix in Table 3.

Table 1. General confusion matrix

		True Class		
		G	B	
Predicted Class	G	g_G	g_B	g
	B	b_G	B_B	b
		n_G	n_B	n

Table 2. Example of a confusion matrix

		True Class		
		G	*B*	
Predicted Class	*G*	600	100	700
	B	150	150	300
		750	250	1000

Table 3. Confusion matrix with a different cutoff

		True Class		
		G	*B*	
Predicted Class	*G*	670	130	800
	B	80	120	200
		750	250	1000

The actual error rate is (130+80)/1000=0.21, which suggests this is a better cutoff. However, the actual expected loss rate is (100*80)+(500*130)/1000=73, which is higher than the expected loss with the other cutoff. So perhaps the other system was superior. This difference between the two ways of trying to decide which system to go for is very common. Sometimes in credit scoring, one will find that the error rate is minimized by classifying everyone as good and so accepting them all. It is a brave---and foolish---credit analyst who suggests this and thus makes himself redundant! One approach that is useful in comparing the difference between the way two credit-scoring systems perform on a holdout sample is to look at the swap sets. This is the group of people in the holdout sample, who are classified differently by the two systems. The swap sets for Table 2 and 3 might look like Table 4.

Table 4. Example of a swap set analysis

	True	
	G	*B*
Predicted *G* by Table 2, *B* by Table 3	50	10
Predicted *B* by Table 2, *G* by Table 3	120	40

This table cannot be calculated from the two confusion matrices, although from these we can see that 100 more in the sample were predicted *B* by Table 2 and *G* by Table 3 than were predicted *G* by Table 2 and *B* by Table 3. There could be in this case a number of people who move a different way to the overall trend. The fact that (50+10+120+40)/1000=0.22 of the population change between the two scorecard suggests that the scorecards are quite different. If one looks only at the swap sets caused by changing the cutoff, then obviously no customers will have their own classifications changed against the movement. Hence one of the rows in the swap sets will be 0.

3 Separation Measures: Mahalanobis Distance and Kolmogorov-Smirnov Statistics

A number of measures used throughout statistics describe how far apart the characteristics of two populations are. If one has a scoring system that gives a score to each member of the population, then one can use these measures to describe how different the scores of the good and the bad are. Thus these approaches can be used only for credit-scoring systems that actually give a score, like the regression approaches or linear programming [7,8,9]. They cannot be used for the credit-scoring systems that group, like classification trees, or where a score is not explicit, like neural networks. Moreover, they describe the general properties of the scorecard and do not depend on which cutoff score is used. This is useful in that these measures give a feel for the robustness of the scorecard if the cutoff score is changed and may be useful in determining what the cutoff score should be. However, when it comes to it, people will want to know how well the scorecard will predict, and to know that one needs to have chosen a specific cutoff score so that one can estimate the error rates and confusion matrices of section 2. For accurate estimates, we should calculate these measures on a holdout sample, which is independent of the development sample on which the scorecard was built. However, often speed and the fact that in many statistical packages it is much easier to calculate the measures on the development sample than on the holdout sample mean that one calculates them first on the development sample. Since they are only indicators of the relative effectiveness of different scorecards, the assumption is made that if one is much better than the other on the development sample, it will remain so on the holdout sample.

The first measure, the Mahalanobis distance, appeared earlier in the discussion on the Fisher approach to using linear regression as a classification method. There we found the linear combination Y of the application variables so that M, the difference between the sample mean of Y for the good and the sample mean Y of the bad, divides by the standard deviation in Y for each group was as large as possible. This M is the Mahalanobis distance and is a measure of by how much the scores of the two groups of the good and the bad differ. Formally, if $n_G(s)$ and $n_B(s)$ are the numbers of good and bad with score s in a sample of n, where there are n_G good and n_B bad, the $p_G(s) = n_G(s)/n_G$ ($p_B(s) = n_B(s)/n_B$) are the probabilities of a good (and bad) having a score s. Then $m_G = \Sigma s p_G(s)$ and $m_B = \Sigma s p_B(s)$ are the mean scores of the good and bad. Let σ_G and σ_B be the standard deviation of the scores of the good and the bad, calculated as

$$\sigma_G^2 = \left(\sum_s s^2 pG(s) - m_G^2 \right)^{\frac{1}{2}} . \sigma_B^2 = \left(\sum_s s^2 pG(s) - m_B^2 \right)^{\frac{1}{2}} \quad (7)$$

Let σ be the pooled standard deviation of the good and the bad from their respective means: it is calculated as follow:

$$\sigma = \left(\frac{n_G \sigma_G^2 + n_B \sigma_G^2}{n} \right)^{\frac{1}{2}} \quad (8)$$

The Mahalanobis distance M is then the difference between the mean score of the two groups, suitably standardized

$$M = {(m_G - m)}/{\sigma} \tag{9}$$

This is indifferent to any linear scaling of the score, and as Figure 1 suggest, one would assume that if a scorecard has a large Mahalanobis distance, it will be a better classifier. In Figure 1, the dotted lines represent possible cutoff scores, and the errors in the figure on the left with the smaller M are much greater than those on the right.

The Mahalanobis distance measures how far apart the means of the good score and the bad score are. The Kolmogorov-Smirnov statistics measures how far apart the distribution functions of the scores of the good and the bad are. Formally, if $P_G(s) = \Sigma_{x \le s} p_G(x)$ and $P_B(s) = \Sigma_{x \le s} p_B(x)$ (or the sums replaced by integrals if the scores are continuous), then the Kolmogorov-Smirnov (KS) statistic is

$$KS = \max_s \left| P_G(s) - P_B(s) \right| \tag{10}$$

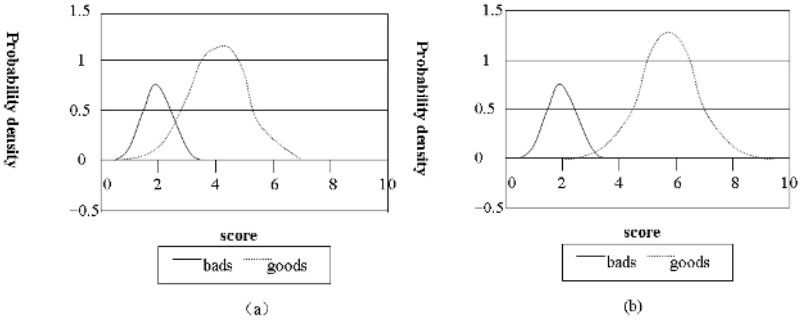

Fig. 1. a) Good and bad similar, b) good and bad different

In Figure 2, the Kolmogorov-Smirnov statistic is the length of the dotted line at the score that maximizes the separation in the distribution function. If the distribution functions are sufficiently regular, then the Kolmogorov-Smirnov distance occurs at the score where the good and bad histograms in Figure 1 cross. The Kolmogorov-Smirnov statistic for an attribute, rather than the score, is used to find the best splits in a classification tree.

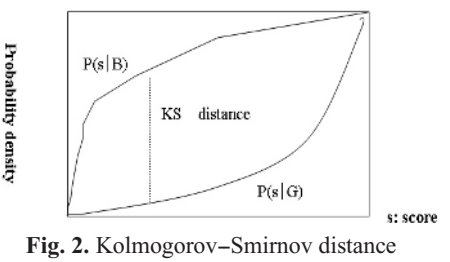

Fig. 2. Kolmogorov–Smirnov distance

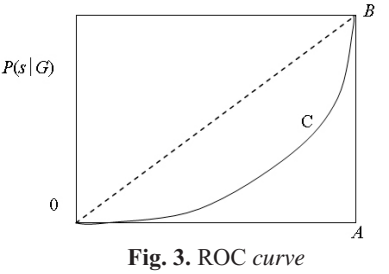

Fig. 3. ROC *curve*

4 Concluding Remarks

Kolmogorov-Smirnov statistics can be displayed, as in Figure 2, by plotting two curves, but it is possible to display the same information on one curve by plotting $P_G(s)$ against $P_B(s)$.The result is a curve as in Figure 3, where each point on the curve represents some score s, and its horizontal distance is $P_B(s)$ and its vertical value is $P_G(s)$. This is the ROC curve, sometimes called the Lorentz diagram. It describes the classification property of the scorecard as the cutoff score varies. The best possible scorecard would have an ROC curve that goes all the way along the horizontal axis before going up the vertical axis. Thus point A would correspond to a score s*, where $P_B(s^*) = 1$ and $P_G(s^*) = 0$; i.e., all of the bad have scores less than s* and none of the good do. An ROC curve along the diagonal $O\ B$ would correspond to one where at every score $P_G(s) = P_B(s)$, so the ratio of good to bad is the same for all score ranges. This is no better than classifying randomly given that one knows the ratio of good to bad in the whole population. Thus the further from the diagonal the ROC curve is, the better the scorecard. If one scorecard has a ROC curve that is always further from the diagonal than the ROC curve of another scorecard, then the first scorecard dominates the second and is a better classifier at all cutoff score. More common is to find ROC curves that cross so one scorecard is a better classifier in one score region and the other is better in the other region.

References

1. Capon, N.: Credit scoring systems: A critical analysis, J .Marketing, 46 (1982) 82-91.
2. Churchill, G.A., Nevin, J.R., Watson, R.R.: The role of credit scoring in the loan decision, Credit World, March (1997) 6-10.
3. Crook, J.N.: Credit constraints and U.S. households, Appl. Financial Economy, 6 (1996) 477-485.
4. Crook, J.N.: Consumer credit and business cycles. In: D.J. Hand and S.D. Jack (eds.): Statistics in Finance. Arnold, London (1998).
5. Edelman, D.B.: Building a model to forecast arrears and provisions. In Proceedings of Credit Scoring and Credit Control VI. Credit Research Center, University of Edinburgh Edinburgh, Scotland (1999).
6. Hand, D.J., Henley, W.E.: Statistical classification methods in consumer credit. J. Roy. Statist. Soc. Ser. A, 160 (1997) 523-541.
7. Haykin, S.: Statistical Aspects of Credit Scoring, PH. D thesis, Open University, Milton Keynes, U. K. (1995).
8. Kou, G., Liu, X., Peng, Y., Shi, Y., Wise, M., Xu, W.: Multiple Criteria Linear Programming to Data Mining: Models, Algorithm Designs and Software Developments. Optimization Methods and Software, 18 (2003) 453-473.
9. Shi, Y., Wise, M., Luo, M. and Lin, Y.: Data mining in credit card portfolio management: a multiple criteria decision making approach, in M. Koksalan and S. Zionts, eds., Multiple Criteria Decision Making in the New Millennium, Springer, Berlin (2001) 427-436.
10. Shi, Y., Peng, Y., Xu, W., Tang, X.: Data Mining via Multiple Criteria Linear Programming: Applications in Credit Card Portfolio Management. International Journal of Information Technology and Decision Making. 1 (2002) 131-151.