Development and Interpretation of a Credit Risk Evaluation Instrument

T. Hillman Willis
College of Business Administration
Louisiana Tech University
Ruston, Louisiana

Banks and other commercial lending institutions frequently use a credit "score card" instrument to evaluate the credit worthiness of potential borrowers. Regulation B of the Code of Federal Regulations (12 CFR 202), effective 1977, stipulates the requirements pertaining to the Credit Opportunity Law. In essence, the regulation prohibits the denial of credit on the basis of age, marital status, location of residence, race, or sex. Further, the regulation requires that every commercial credit lending institution develop an objective credit scoring system on a statistically derived set of criteria stemming from empirical data pertaining to the credit history of loan applicants of that particular institution.

Therefore, the end product of a credit scoring system is a score card instrument that furnishes a numerical evaluation of the criteria for assessing credit worthiness. The purpose of this paper is to explain a statistical procedure for developing a score card and how the use of a score card can improve the loan approval decision. The improvement is demonstrated using a Bayesian approach.

The use of a point allocation system to measure the risk of a loan applicant has been used for several years and has been much discussed ([2], [7], [9], [10], [12]). A point score system can lend a certain amount of objectivity to what would otherwise be a personal judgment decision. Many businesses use a point system that is simply an assignment of quantitative values for various personal data traits that are based on the subjective assessment of a "credit expert." These point systems could be considered to be based on empirical data since the "expert" is using his or her experience with the credit history of borrowers. However, the development of a credit scoring system using statistical analysis and validation as required by Regulation B, is not very common in the lending industry.

Durand [5] wrote the first significant and comprehensive work on credit scoring. His system was based on a univariate test for significance which is now outdated because of the availability of the computer to perform more sophisticated analyses. Management Decision Systems, Inc. of Georgia [4] presents a more current procedure for developing a credit scoring system, but crucial details are left to the reader's imagination. This paper attempts to shed some

Journal of Business Strategies, Volume 5, Number 1 (Spring 1988)

light on the gray areas that are usually glossed over by the industry due to their confidential nature.

Credit Risk Variables

The first step involves the selection of the variables included on the score card. A typical credit score card would contain several variables with a certain number of points allocated for each separate classification for each variable. Such variables typically include years at present address, years at present job, occupation classification, annual salary, monthly housing payment, number of people in household, housing classification (own, rent, other), etc.. For example, the variable, time at present address, may be allocated 12 points for six months or less, 16 points for over six months to one year, 23 points for over one year to two years, etc.

The points allocated to each category are not arbitrary values. They are computed from the discriminant function as explained in the next section. A higher total score reflects a better credit risk. Therefore, the points computed for each characteristic represent the relative importance of a particular categorical response to account for a good credit risk. For example, if living at the present address for over ten years is worth 32 points as compared to only 16 points for one year residence, a person with the longer time at the same residence is deemed the better risk by a factor of 2 to 1.

A typical loan application normally includes an extensive list of variables. On the basis of the responses provided on the loan application, certain variables are scored and totaled. If the total exceeds a statistically derived cutoff value the loan would qualify for approval; otherwise, it would not. An alternative procedure would be to include another category, one for which a minimum and a maximum score is established. Any score between these limits would be sent to the loan supervisor for further deliberation.

An example of a credit score card developed for a medium–sized bank located in the midsouth is shown in Appendix A. Suppose an applicant has the following characteristics: (1) located at the present address for 2 years; (2) buying the residence; (3) monthly payment is $600; (4) four people in the household; (5) working at the present job for 5 years; (6) office worker; (7) no savings account; (8) paid out a bank loan; and (9) monthly income of $1,800. Using the score card in Appendix A, the point total is 240, which means that the applicant would automatically qualify for approval if the suggested cutoff guide is used.

The process of selecting the particular variables to be used and allocating the point values to the various intervals is the critical part of the process. First, an appropriate data base must be established. A sufficient history of loan activity must be available to establish a block of "good" and "bad" loans. It is desirable to have several hundred loans of both types and preferably on computer tape or disk for processing convenience.

The size sample of "good" and "bad" loans that is sufficient for statistical analysis depends upon the number of credit characteristics evaluated. However, a sample of 100 loans of each type would be sufficient in most cases. Actually, two samples are needed. An analysis sample is needed to construct the score card. A second sample, called the "holdout" or "validation" sample, is used to validate both the discriminant function and the score card. This process is discussed later.

Of course, the larger the sample the better. The lending institution illustrated in this paper had a loan population of sufficient size to permit a random sample of 500 good and 500 bad loans. Half of these were used for analysis and half for validation. To get a sample this large requires that the loan application be standardized and used for a substantial period of time. In fact, the loan application should be used long enough for the loans to be either paid out (or a certain history of good credit to be built up) or defaulted (or excessive late payments to be recorded to result in a "bad" loan).

The definition of a "bad" loan may vary among loan institutions, but it is up to the lender to make that determination. In most instances a loan default is not all that is required to result in a "bad" loan. A record of one payment over 90 days late, or two occasions of over 60 days late, or three or more occasions of being over 30 days late may be considered "bad." There appears to be no standard definition.

The next step is to determine which variables are potential predictors of credit worthiness. This can be done with various procedures, the most popular of which is the chi-square test. The chi-square distribution is used to perform a contingency table test. The table consists of the number of loans that were found to be good or bad and classified according to an appropriate variable scale. An example of a chi-square contingency table is shown in Table 1.

Table 1

Contingency Table for Housing Classification Variable (number of loans)

Housing Classification	Good Loan	Bad Loan
Buying or Own	351	286
Rent	92	121
Other	57	93
Total	500	500

$\chi^2 = 19.22$ with 2 degrees of freedom

p-value < 0.005

Table 1 shows that more of the good loan applicants are buying or own their residence. On the other hand, more of the bad loan applicants either rent or fit some other classification, such as live with relatives. The p-value less than

0.005 indicates that this particular variable is very useful in predicting whether the applicant is a good credit risk.

If the variable is dichotomous, such as the presence of a current checking account or savings account, or discrete, such as the housing classification as shown in Table 1, the classes are obvious. But if the variable is continuous, the classes are not as obvious. For example, annual income, monthly housing payment, time at present job, and time at present address are variables that require class intervals to be specified.

To determine the appropriate intervals, a sorted listing of the data for each variable is helpful. From this, the range and any natural groupings of the data can be identified. Then it is a matter of testing various group intervals to determine the particular classifications that give the largest chi–square value. Of course this extra effort is only justified after the variable has been identified as significant in distinguishing between good and bad loans. It is also possible to use a series of discriminant analyses to locate the most discriminating interval breaks, but it is recommended only when computer time is not a constraint.

Those variables that have significant relationships are selected for further consideration. A significance level should be selected for this decision. Although significance levels of 0.01 or 0.05 are common, the researcher used a cutoff of a p–value of 0.10 or less. This figure would allow elimination of those variables showing no significant relationship and those being marginal would be tried because they could prove useful in combination with other variables.

Linear Discriminant Analysis

Those variables that have statistically significant relationships to credit worthiness are submitted to a multiple linear discriminant analysis program. This procedure is well documented ([1], [3], [6], [8]) and is available on several computer packages. The most common packages are SAS and SPSS. A loan application may have thirty or forty data values that could be useful in predicting credit worthiness. After employing chi–square tests, as many as twenty variables may appear to be significant. The discriminant analysis program is then run on these variables to determine which variables should be included on the score card. As evidenced in Appendix A, the typical score card usually contains ten or fewer evaluation criteria.

A discriminant analysis is appropriate when the population of interest consists of two distinct groups–good and bad credit risks. An F–ratio is calculated from the analysis sample to measure whether the two loan type groups have been significantly separated on the basis of the sample loans. The necessary assumptions for the multiple linear discriminant function are equal dispersion matrices for each group and a multivariate normal distribution of the population.

Determining the best set of explanatory variables is a decision based on the results of the discriminant analysis. This is accomplished by evaluating the F statistic, the standard error of estimate, the multiple coefficient of determination, and the partial correlation coefficients. The discriminant analysis accomplishes two things. First, it identifies those variables having the greatest explanatory value and, second, it provides discriminant linear coefficients which are used to calculate the points for the score card.

The procedure for determining the actual points uses the ratio of good-to-bad loans for each class interval for each variable. Each variable will have an interval either established statistically for the data (e.g. monthly income–\$0–499, 500–999, 1000–1499, etc.) or the interval will be natural (e.g. housing classification–renting, buying, or other). The discriminant analysis coefficients for each variable are then multiplied by the ratio of good–to–bad loans that applies to each interval. The resulting values are the scores for each category.

Validation of the discriminant function is accomplished using the holdout sample to construct a confusion (cross–classification) matrix and computing the F statistic level and Wilk's lambda to determine the significance of these characteristics. The confusion matrix gives the percentage of good and bad loans that are correctly or incorrectly classified and provides insight regarding the performance of the discriminant equation.

Score Card Validation

Another important use of the "holdout" sample is to use those loans to test the effectiveness of the score card and validate the score card. This is done by first computing a total score for each loan known to be good or bad. When these total scores are ranked, it is possible to estimate the percentage of good and bad loans that would be accepted at various cutoff scores. This allows the lending institution to predict the effect of selecting a particular cutoff score. A typical validation sample prediction table is given in Table 2.

To illustrate what the percentages in Table 2 represent, suppose a credit score of 110 is considered. The table shows that no loans (good or bad) scored below this value. For those loans scoring 170 or lower, 89 percent were good loans and 58 were bad. In fact, none of the loans that turned out to be bad scored higher than 270 whereas the good loans scored as high as 310. Thus, as the score gets higher, the percentage of bad loans drops faster than the percentage of good loans.

As shown in Table 2, the cutoff score resulting in the greatest differentiation between a good and bad loan is 210. In other words, if a loan scoring below 210 is rejected as too risky and one scoring 210 or above is accepted, then the result is to accept two–thirds of the potentially good loans and only one–fourth of the potentially bad loans. In actual practice, a minimum and a maximum cutoff are usually established. For example, a loan scoring below 170 is automatically rejected and one scoring 240 or above is automatically approved. If a loan scores between these two limits, then the loan supervisor can make the final decision.

One of the difficulties in developing a credit scoring system is that the sample should be representative of the entire population which includes accepted and rejected applicants. There are several ways to deal with the problem of including rejected applicants. One is to simply ignore them and this is often the case. Another is to assume that the loan officer made the correct decision and, therefore, consider a rejected applicant a bad loan and include them in the "bad" population. A more valid approach is to score each rejected loan as either a good or a bad loan and use them to augment the population. In that case it will be necessary to take a new sample from this adjusted population and rerun the discriminant analysis and recompute the points for the score card. Experience has shown that including the rejected loans only slightly affects the results.

Table 2

Validation Sample Prediction Table

Cutoff Score	Percentage of "Good" Loans Accepted	Percentage of "Bad" Loans Accepted	Percentage Difference
110	100	100	0
130	99	94	6
150	97	82	15
170	89	58	31
190	80	39	41
210	67	25	42*
230	52	15	37
250	32	6	26
270	15	1	14
290	6	0	6
310	2	0	2
330	0	0	0

* = Largest percentage difference in sample

Bayesian Interpretation

The next step is to determine the effectiveness of the credit scoring instrument. A Bayesian procedure can be used to predict the performance of the score card. Assume that the minimum cutoff score of 170 is set. As shown in Table 2, this score means that 89 percent of the good loans and 58 percent of the bad loans are accepted. Next, suppose that the loan history for this particular institution prior to the use of a score card is as follows:

% of Total applicants granted and good	=	49%
% of Total applicants granted and bad	=	9%
% of Total applicants rejected	=	42%
		100%

Notice that the population should also include the rejected loans. Fifty-eight percent were approved and 42 percent were rejected. Of the granted loans in the past, 84 percent (0.49/0.58 = 0.84) were good and 16 percent (0.09/0.58 = 0.16) were bad. Multiplying the historical percentages by the probabilities that would apply when using the credit scoring model with a cutoff of 170, gives the following joint probability table.

	Good	Bad	Reject	Total
Accepted by the model	.44	.05	.24	.73
Turned down by the model	.05	.04	.18	.27
	.49	.09	.42	1.00

The table shows the breakdown of how each type of loan would be treated by the scoring system. For example, of the 49 percent "good" loans, 44 percent (0.49 × 0.89 = 0.44) would be accepted by the model and 5 percent (0.49 - 0.44 = 0.05) would be turned down. Of the 9 percent "bad" loans, 5 percent (0.09 × 0.58 = 0.05) would be accepted by the model and of the 42 percent rejected loans, 24 percent (0.42 × 0.58 = 0.24) would be accepted by the model. Overall, the credit scoring model would accept 73 percent of the loan applicants and turn down 27 percent.

The Bayes formula used to revise the probability of a good loan is:

$$P(A_1|A.C.) = \frac{P(A.C.|A_1) \times P(A_1)}{P(A.C.|A_1) \times P(A_1) + P(A.C.|A_2) \times P(A_2)} \qquad (1)$$

where $A.C.$ = "above cutoff," A_1 = good loan, and A_2 = bad loan.

The Bayes revision of the prior probabilities using the score card gives the following projection of the percentages of good and bad accounts that would be accepted by the model.

Percentage of Accepted that are Good = 90% (0.44/(0.44 + 0.05))

Percentage of Accepted that are Bad = 10% (0.05/(0.05 + 0.44))

These probabilities can be summarized as follows:

					P(A.C.	
	P(A_i)	×	P(A.C.|A_i)	=	and A_i)	P(A_i|A.C.)
A_1 - "Good"	.49	×	.89	=	.44	.44/.49 = .90
A_2 - "Bad"	.09	×	.58	=	.05	.05/.49 = .10
					.49	1.00

Based on these figures, the forecasted percentages for a cutoff score of 170 can be calculated.

Percent above cutoff and " good" = .90 × .73 = .66

Percent above cutoff and "bad" = .10 × .73 = .07

The performance of the credit scoring instrument can be illustrated by contrasting the probabilities associated with using a score card with those associated with not using a score card. This is shown in Table 3.

The result is that the scoring system is estimated to increase the percentage of good loans that are accepted from 49 to 66 percent which is a 17 percent improvement. The percentage of bad loans that are accepted is decreased by two percent. In addition, the percentage of loans that need to be rejected is reduced by 15 percent.

Table 3

Performance Comparison
Using a Credit Scoring Instrument

	Score Card (Cutoff = 170)	No Score Card
% of total applicants accepted and "good"	66%	49%
% of total applicants accepted and "bad"	7%	9%
% of total applicants "rejected"	27%	42%
	100%	100%

Remember that this example is intended to merely illustrate the performance of the scoring instrument. Experience has been that greater improvements in decision making are possible depending on the particular loan history of the lending institution.

Conclusions

The use of an empirically derived credit scoring system has two advantages. One of these is that it enables the lending institution to be in compliance with federal regulations and furnishes a valid defense in case of litigation. Second, when the score card is based on sound statistical practices, its use can often improve the lending decisions of the institution as compared with strictly subjective assessments. A credit scoring instrument can play an important role in the overall lending strategy of financial institutions.

References

1. Berenson, M. L., D. M. Levine, and M. Goldstein. *Intermediate Statistical Methods and Applications: A Computer Package Approach.* Englewood Cliffs, NJ: Prentice Hall, Inc. (1983).

2. Bierman, H. Jr. and Hausman, W. H. "The Credit Granting Decision." *Management Science*, Vol. 16 (April 1970), pp. B519–B532.

3. Cooley, W. W. and Lohnes, P. R. *Multivariate Data Analysis*, New York, NY: John Wiley and Sons, Inc. (1971)

4. *Credit Scoring Systems: A Detailed Analysis.* Atlanta, GA: Management Decision Systems, Inc. (1977).

5. Durand, D. *Risk Elements in Consumer Installment Financing.* New York, NY: National Bureau of Economic Research (1941).

6. Goldstein, M. and Dillon, W. R. *Discrete Discriminant Analysis.* New York, NY: John Wiley and Sons, Inc. (1978).

7. Johnson, N. "How Point Scoring Can Do More Than Help Make Loan Decisions." *Banking*, Vol. 62 (August 1971), pp. 36–42.

8. Morrison, D. G. "On the Interpretation of Discriminant Analysis." *Journal of Marketing Research*, Vol. 6 (May 1969), pp. 156–163.

9. Myers, J. H. "Predicting Credit Risk with a Numerical Scoring System." *Journal of Applied Psychology*, Vol. 47 (October 1963), pp. 348–352.

10. Orgler, Y. E. "Evaluation of Bank Consumer Loans with Credit Scoring Models." *Journal of Bank Research*, Vol. 2 (Spring 1971), pp. 31–37.

11. Rock, A. "Sure Ways to Score with Lenders." *Money*, September 1984, pp. 121–126.

12. Roy, H. J. H. and Lewis, E. M. "Credit Scoring as a Management Tool." *Consumer Credit Leader*, Vol. 1 (November 1971), pp. 10–13.

Appendix A
Credit Score Card Example

Step 1: After completion of the loan application in its entirety, score each of the characteristics below and then total the point values.

Evaluation Characteristics	Qualities					
	Point Values					
Time at Present Address	6 mos. or less	7 mos. to 1 yr.	1 yr. 1 mo. to 2 yrs.	2 yrs. 1 mo. to 6 yrs.	6 yrs. 1 mo. to 10 yrs.	Over 10 years
	12	16	23	28	38	32
Housing Classification	Rents	Buys or owns	Other			
	10	32	14			
Monthly Rent or Payment	$0	$1 - 200	$201 - 300	$301 - 400	$401 and over	Owns free and clear
	5	6	8	14	19	10
Number in Household	1	2	3	4	5 or more	
	38	52	41	36	26	
Time at Present Job	6 mos. or less	7 mos. to 2 yrs.	2 yrs 1 mo. to 4 yrs.	4 yrs. 1 mo. to 6 yrs.	Over 6 yrs. and retired	
	4	6	8	9	13	
Occupation	Prof./Exec. Mgr./Ret.	Sales	Semi-prof. Office/Staff	Lab./Service Prod./Driver	Other	
	38	35	25	14	16	
Savings Account	None	Yes				
	22	43				
Paid Out Bank Loan	None	Yes				
	14	60				
Total Monthly Application Income	$0 - 750	$751 - 1000	$1001 - 1500	$1501 - 2000	$2001 - 3000	$3001 and over
	7	9	10	14	21	22

Step 2: Evaluate the total using the following guide:

169 or less – Reject.
170 to 239 – Go to next step.
240 or higher – Qualifies for approval