The Development of Numerical Credit Evaluation Systems

James H. Myers; Edward W. Forgy

Journal of the American Statistical Association, Volume 58, Issue 303 (Sep., 1963), 799-806.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28196309%2958%3A303%3C799%3ATDONCE%3E2.0.CO%3B2-X

THE DEVELOPMENT OF NUMERICAL CREDIT EVALUATION SYSTEMS

JAMES H. MYERS
University of Southern California
AND
EDWARD W. FORGY
University of California at Los Angeles

Several discriminant and multiple regression analyses were performed on retail credit application data to develop a numerical scoring system for predicting credit risk in a finance company. Results showed that equal weights for all significantly predictive items were as effective as weights from the more sophisticated techniques of discriminant analysis and "stepwise multiple regression." However, a variation of the basic discriminant analysis produced a better separation of groups at the lower score levels, where more potential losses could be eliminated with a minimum cost of potentially good accounts.

1. INTRODUCTION

THE problem of determining credit risk has been with the businessman since he first consummated a business transaction without receiving immediate payment for his goods or services. His ability to extend credit profitably has depended simply upon his own good judgment, based upon experience gained through previous credit decisions, either his own or those of business associates.

In recent years, attempts have been made to assist the businessman by providing quantitative rating systems for use in evaluating applications for credit at the retail level. Such rating systems assign numerical weights to certain items of personal history information shown on the credit application form (e.g., 10 points for married persons vs. 2 points for unmarried; 15 points if home is owned vs. −3 points if rented). Points from a number of such items are added together to produce a total score for each applicant, and this score is used as a measure of payment potential—the higher the score, the more likely the applicant possesses those personal and family characteristics which indicate a disposition toward prompt payment of credit obligations.

An approach of this sort has been all the more necessary in recent years because of the mushrooming of "credit psychology" among large segments of the consumer public. The resulting increase in demand has taxed credit departments of many larger companies beyond their capacities to train and maintain an adequate staff of experienced credit evaluators, and thus increasing amounts of retail credit are being granted by younger and less experienced individuals. In such a situation, numerical rating systems have much to offer, provided they can predict credit risk with sufficient accuracy.

Previous Studies

The earliest published study dealing with the development of a numerical rating system was that by Durand [1] in 1941, under the sponsorship of the

National Bureau of Economic Research. Hundreds of both good and bad personal loan accounts were analyzed from the files of commercial banks, personal finance companies, industrial banking companies, automobile finance companies, and appliance finance companies. Employing discriminant analysis, Durand developed several different weighting systems for these accounts. Results showed good prediction of credit repayment in 20 commercial banks and in 9 industrial banking companies (in aggregates, not as individual organizations). It is not certain, however, that these results were checked out on new samples to determine possible "shrinkage" of effectiveness in actual use.

A later study by Wolbers [4] in 1949 was done in one branch of a nationwide department store chain. Developing scoring weights on one sample and applying them to a second sample, he showed that credit losses could be reduced approximately 7 per cent, with negligible losses in the volume of good business. In 1957, Myers and Cordner [3] studied credit accounts in one branch of a Los Angeles chain dealing in personal loans. Results showed that approximately 6 per cent of the losses from this branch could be eliminated with *no loss* in the volume of good business; approximately 24 per cent of losses could be eliminated if 3 per cent of good business volume could be sacrificed; and approximately 50 per cent of losses could be eliminated at the cost of only 7 per cent of good business volume. Another study [2], conducted for an automobile dealer, found that approximately 20 per cent of credit losses could be eliminated at the cost of only 1 per cent or so of good business.

While results from these studies have differed, all have shown that a properly constructed numerical rating system can offer at least *some* degree of improvement over the purely subjective or judgmental approach to evaluating credit. In some cases, the amount of this improvement has been substantial.

In spite of these results, numerical rating systems are not in widespread use today. There may be several reasons for this: 1) a natural reluctance on the part of experienced credit executives to abandon the time honored "judgmental" approach in favor of newer and relatively untested quantitative rating methods; 2) the inability of statisticians to develop "fool-proof" rating systems which consistently identify poor credit risks accurately enough to result in substantial net savings by the credit operation; 3) the difficulties involved in utilizing any such effective rating system in the operating situation; and 4) unwillingness on the part of statistical consultants to invade the domain of the credit manager and do the selling job necessary to transform such an idea into a successful and useful operating tool.

2. SCOPE OF PRESENT STUDY

The study reported here was done in the central office of the Universal Finance Company, Los Angeles, California. This company purchases conditional sales contracts (i.e., title to the property does not pass to the buyer until all payments are made) on mobile homes, primarily in the Western portion of the U. S. Contracts are purchased from mobile home dealers through a numbet of company branch offices located in various cities in the West.

The Company had developed by itself a "deal grading" numerical scoring

system[1] based upon the pooled judgment of experienced credit evaluators. This system had been in use a short while at the time the present study was undertaken. The purpose of the study reported here was to develop a new scoring system to replace the existing "deal grading" scoring system. While "deal grading" offered some improvement over the former purely subjective evaluation of purchase applications, it was felt that a more intensive statistical study of accounts might produce a still better scoring system.

3. METHOD

To develop a new scoring system, 600 Universal Finance accounts were randomly selected, as follows:

Initial Sample	Hold-out Sample	Definition
150	150	Good accounts (Open) Opened between January, 1959 and October, 1960. No more than 2 or 3 late payments.
150	150	Repossessions (Closed) Opened in 1959 or later. Made less than 18 payments. Repossessed in 1960 or 1961.
300	300	

Item selection and scoring weight analyses were done on the "initial sample" of 300 cases (150 good, 150 repossessions). The various scoring systems developed were then applied to the "hold-out" sample of 300 cases to produce an *unbiased* estimate of what each scoring system could be expected to do in actual operation. This procedure eliminated any systematic capitalizing upon chance fluctuations in the particular sample of accounts used in the initial analysis which would have produced a spuriously high degree of discrimination between good and bad groups.

For each of the cases chosen, a large number of items were coded relating to the *personal history of the applicant* (e.g., age, previous credit record, time on present job, number of children, type of bank account, etc.) and the *transaction itself* (per cent down payment, width of mobile home, etc.). For a complete list of all items coded, see Table 1.

To determine whether or not a particular item would predict payment vs. non-payment of an account, the responses to that item were compared between the two groups: 1) good accounts and 2) repossessions. For example, it was found that only 13 per cent of the good account group had no bank account, while 31 per cent of the bad (repossession) group had no bank account. Thus, this item could be said to be predictive of future credit payment potential: applicants with no bank account would be less likely to pay than those with bank accounts.

[1] Consisting of the following items: per cent down payment, unpaid balance of transaction, time at present job, income, and an over-all quality rating.

TABLE 1. ITEMS CODED FOR ANALYSIS AND PREDICTIVE ITEMS

1.	P-F	Age of principal
2.	_____	Married?
3.	_____	Age difference between man and wife
4.	_____	No. children
5.	_____	No. "other" dependents
6.	P-F	Now have phone?
7.	P	Will have phone at new address?
8.	_____	Change of city or state from previous address?
9.	P-F	New parking address (Trailer park vs. private property)
10.	P-F	Mailing address different than new parking address?
11.	_____	Trailer for pleasure?
12.	_____	Occupation—Principal wage earner (P.W.E.)
13.	_____	Both work?
14.	_____	Total monthly income
15.	_____	Work phone? (P.W.E.)
16.	P-F	Time at present job
17.	_____	Time at present & past jobs combined
18.	_____	Trade union?
19.	_____	Other income?
20.	P-F	Bank account?
21.	P-F	Previous trailer cash?
22.	P-F	Auto clear?
23.	_____	Auto make
24.	_____	Auto year
25.	_____	Location of relatives
26.	P	Own real estate?
27.	_____	Present monthly installment payments as % of #14
28.	P	Previous high credit amount
29.	_____	# Installment references
30.	P-F	# Credit references UNsatisfactory
31.	_____	# Credit references Satisfactory
32.	P-F	# Repos., bankruptcy, suits
33.	P-F	Width of trailer
34.	P	Length of trailer
35.	_____	Expandable?
36.	P-F	Form of down payment
37.	P	% down payment
38.	P-F	Unpaid balance*
39.	P-F	Buying new or used?
40.	P-F	Unpaid balance scaled
41.	P	Term of contract

P = Predictive item in the Initial Sample.
F = Item included in final scoring system.
* Ratio of unpaid balance on present deal to highest previous credit balance (excluding home mortgage).

All coded items (approximately 40) were analyzed in this way, and 21 were found to be predictive of account payment at the .05 significance level or better. The predictive items are indicated by a "*P*" on the line in front of the item in Table 1.

Several of these items were not continuous in nature or even ordered, but rather involved purely nominal categories (e.g., type of back account: checking only, savings only, loan only, combinations of these). Before further analysis could be undertaken, it was necessary to "scale" these items into a "quantified" form for each item so that a direction from best to worst would exist for purposes of intercorrelations and other statistical operations. Scaling was accomplished by assigning high values to favorable categories (characteristic of good accounts), low values to unfavorable categories (characteristic of delinquent accounts).

When all items were in quantified form, the proper selection of items and weights were determined in a number of ways, as described below. Generally, the analyses were done by a series of statistical analysis programs written for the IBM 7090 computer at the Western Data Processing Center, U.C.L.A., by the Division of Biostatistics, School of Medicine, U.C.L.A.

It might be argued that the proper approach would be to develop weights based on *all* 40+items which were coded. However, machine storage limitations

restricted the total number of variables which could be analyzed at one time to 25. It is thus possible that one or more "suppressor variables" (i.e., variables correlating zero with the criterion but positively with one or more predictive variables) were overlooked. With variables of the type used in this study, however, this did not seem too likely.

Alternative Scoring Systems

Rather than using only *one* approach to developing proper weights for the predictive items, it was decided to try several alternative methods. The basic approach was discriminant analysis. However, several modifications of this procedure were used as described below.

Because of the practical importance of identifying as many bad credit risks as possible with a very small cost of good risks, a novel approach was used to construct a scoring system. The objective of this (System IV, below) was to improve discrimination specifically at *lower* score ranges, rather than over-all separation of group means (which is maximized by discriminant analysis).

System No.	Description
I	Conventional discriminant analysis.
II	Stepwise regression. This program performs a series of multiple linear regressions, starting with the most predictive single variable and adding, one step at a time, the one "new" variable which makes the greatest improvement in goodness of fit. This continues until a specified level of significance of improvement by the last new variable fails to be reached.
III	Equal weights for all predictive variables used.
IV	Discriminant analysis weights based upon selected subsamples of cases. The *initial* sample of 300 was scored on System I, and good and bad risks were ranked separately by this score. Then the lowest 250 cases (125 good and 125 bad) were used to compute a new set of discriminant weights. The same was done for the *lowest* 200, 150, 100 and 50 cases.

4. RESULTS

Table 2 shows the results from applying the various alternative scoring systems developed on the initial sample to the 300 "hold-out" cases, representing an unbiased estimate of the relative effectiveness of these scoring systems. Effectiveness of the scoring system was based upon two comparisons:

1) Correlation (point biserial) of actual and predicted scores for each scoring system. Results are in Column 3 of Table 2.
2) Number of bad (repossession) accounts which would be eliminated at a cost of the indicated number of good accounts. For example, the first data line of Table 2 indicates that by using discriminant analysis weights (Weighting System I) for all 21 predictive items, 10 (out of 150) bad cases can be eliminated without losing *any* good cases, 41 bad cases can be eliminated at a cost of five (out of 150) good cases, etc.

Different numbers of items were used in many of the alternative weighting

TABLE 2. PREDICTIVE EFFECTIVENESS OF VARIOUS WEIGHTING SYSTEMS ON HOLD-OUT SAMPLE

(1) Weighting System	(2) No. Items	(3) r_{pbi}	No. bad cases eliminated at cost of indicated no. of good cases				
			0	1	5	10	20
I. Discriminant Analysis	21	.424	10	11	41	59	67
II. Stepwise Regression	15*	.414	9	12	39	53	68
	12*	.402	15	18	39	54	68
III. Equal Weight	21	.483	13	19	35	52	76
	15*	.468	9	13	40	58	77
IV. Discriminant Analysis, selected cases							
($N=250$)	21	.399	17	18	44	52	62
($N=200$)	21	.402	17	19	37	58	64
($N=150$)	21	.420	24	26	42	52	64
($N=100$)	21	.378	21	21	37	40	61
($N=50$)	21	.428	34	35	43	56	67

* "Best" items selected by stepwise regression. $F \geq 1.0$ for 15 items, $F \geq 2.0$ for 12 items.

systems developed. Column 2 of Table 2 shows the number of items used for each system. For example, the 15 items line for Weighting System II refers to the scoring system developed by "stepwise regression" for the "best" 15 items, and the results obtained from applying this system to the same 15 items on the "hold-out" group.

Comparisons between Weighting Systems

Perhaps the most surprising result in Table 2 is the superior discrimination power (based upon the correlation criterion) of Weighting System III, equal weights for all 21 items. The point biserial correlation of .483 is higher than that for the discriminant analysis (System I) using all items ($r=.424$). The reason for this is not immediately clear. It may be simply due to an artifact of the particular data or samples selected for this study, or it might reflect the failure of the data to meet the assumptions underlying the development of the other techniques. This kind of result will be watched for in future studies.

The most encouraging results were those emerging from Weighting System IV. This system was specifically developed to improve discrimination power *at the lower score levels*, so that the maximum number of potentially bad accounts could be eliminated with the smallest possible loss in profitable accounts. This is in contrast to the basic discriminant analysis, which develops weights for the maximum over-all separation of *means* of the two groups.

It is also of interest to note that prediction is nearly as effective with fewer variables as with the full 21. A comparison of Systems I and II, for example, indicates that the 12 "best" variables will predict almost as well ($r=.402$) as

the full 21 ($r = .424$). This is obviously of real importance in the operating situation, where the time to produce a numerical score for each applicant must be kept to a minimum, particularly in large-volume credit operations.

5. DISCUSSION

Final Scoring System

The above results were discussed with Universal Finance management, and the decision was made to adopt the 15-item System II as the final scoring system. (In making the decision, consideration also was given to the "face validity" of the items selected by the various scoring systems and to the weights assigned. System IV, in particular, included some weights which would be seen as bizarre by operating personnel.)

While the complete scoring for all items cannot be revealed here, the following item will serve to illustrate the general approach:

Bank Account	Score
None	0
Other (checking OR savings alone, etc.)	2
Checking AND savings	4

A careful comparison showed that simplified weights of the type shown here work as well as more cumbersome two or three digit weights.

Table 3 shows results from applying the 15-item System II weights to the "hold-out" sample. It can be seen, for example, that the rejection of applicants scoring 19 or below will eliminate 13 per cent of the bad accounts at a cost of less than 1 per cent of good accounts; rejecting at a score of 22 or below will eliminate 26 per cent of potential future losses at a cost of 3 per cent of good accounts, etc. With such information, management can set a cut-off score which is in line with profit-loss relationships on good vs. poor business and with company operating policy. Changes in cut-off scores can easily be made to reflect changing economic conditions, changes in operating policy, etc.

From a technical standpoint, the results of this study are encouraging in that they suggest there may be ways in which the basic discriminant analysis approach can be modified to improve its effectiveness, particularly in the case of discrimination at the lower score levels, which are of the greatest practical importance in retail credit evaluation. Indeed, there may well be additional modifications, other than those used in this study, that produce even better results at these lower score levels. Attempts will be made in future studies to develop additional tools along this line.

TABLE 3. SCORING SYSTEM RESULTS: WEIGHTING SYSTEM II (15 ITEMS) APPLIED TO HOLDOUT SAMPLE

If refuse applicants scoring at or below:	Will eliminate:	
	% Bad (Repossessions)	% Good (Open)
19	13%	Less than 1%
20	19	1
21	23	2
22	26	3
23	34	6
24	40	7
25	42	10
26	46	14
27	49	20
28	59	26
29	65	34
30	71	43
31	79	50
32	85	55
33	87	63
34	89	68
35	91	74
36	95	82
37	98	86
38	99	91
39	99	94
40	99	97
41	100	98
42	100	98
43	100	98
44	100	99
45	100	100

(Note: Scores can range from 0 to 51.)

REFERENCES

[1] Durand, David, "Risk Elements in Consumer Installment Financing," (Study #8) National Bureau of Economic Research, New York, 1941.

[2] McGrath, James J., "Improving Credit Evaluation with a Weighted Application Blank," *Journal of Applied Psychology*, 44, (1960) 325–8.

[3] Myers, J. H., and Cordner, Warren, "Increase credit operation profits," *The Credit World*, February 1957, 12–3.

[4] Wolbers, H. L., "The Use of the Biographical Data Blank in Predicting Good and Potentially Poor Credit Risks," Unpublished M. A. Thesis, University of Southern California, 1949.