MANAGEMENT SCIENCE
Vol. 31, No. 2, February 1985
Printed in U.S.A.

A ROBUST CREDIT SCREENING MODEL USING CATEGORICAL DATA*

PETER KOLESAR AND JANET L. SHOWERS
Graduate School of Business, Columbia University, New York, New York 10027
Salomon Brothers Inc, 1 New York Plaza, New York, New York 10004

Motivated by an application in a public utility, the credit screening problem is re-examined from a decision theoretic viewpoint. The relationships between several alternative problem formulations are explored, and compared to the classical linear discriminant analysis (LDA) approach. Several mathematical programming based solution methods are proposed when the data are binary, and an efficient algorithm is developed for the case when the screening function must also have binary weights. Actual results of both the mathematical programming and LDA methods are presented and compared. The resulting mathematical programming rules are effective, robust, and flexible to administer. Practical advantages of the resulting "n out of N" type rules are discussed. These screening rules have been widely implemented by a major public utility and have resulted in substantial benefits to the utility and to the public. (FINANCE; INDUSTRIES, COMMUNICATIONS; STATISTICS; DECISION ANALYSIS; INTEGER PROGRAMMING—APPLICATIONS)

1. Introduction

Many institutions that provide mass credit use statistical screening procedures that translate a background profile sketching the applicant's credit history into a numerical credit score. Applicants with a high enough score pass the screen and are granted credit. In the early 1940's, David Durand (1941) proposed the linear discriminant analysis (LDA) functions that had been developed earlier by R. A. Fisher (1936) for such credit screening. In an important paper Myers and Forgy (1963) reported an application to the financing of mobile homes. There followed a large literature on applications of *credit scoring*—usually using LDA. Orgler (1975) summarizes much of the work. The basic statistical theory is contained in the standard references by Morrison (1976) and Anderson (1958).

Despite the advantages of a well-developed theory and relative computational simplicity the LDA approach has several potential practical disadvantages that led us to consider the problem afresh. First, to facilitate implementation we sought even simpler rules than the linear continuous functions LDA creates. Second, the structure of LDA rules sometimes defies common sense, and can lead to decisions that might be particularly difficult to defend in a public forum. (See Capon 1982 for an extensive discussion of such problems.) Third, LDA produces solutions that are optimal for a particular decision problem when the variables are continuous and have a special multivariate normal distribution. Unfortunately, these conditions don't always hold in practice and the problem optimized by linear discriminant functions is not itself of direct interest in credit screening.

In contrast, this paper takes a simple decision theoretic view of the credit screening problem that focuses on those situations where the data are (binary) categorical variables. We use mathematical programming to solve the problem and our method produces flexible credit screens that are robust and very easy to implement. In most of our work we restrict ourselves to scoring rules that simply count the number of "positive" responses on the credit questionnaire. These "n out of N" rules make

*Accepted by Paul Gray; received February 18, 1984. This paper has been with the authors 4 months for 1 revision.

0025-1909/85/3102/0123$01.25

implementation particularly easy. Remarkably, we find that our procedure loses little in performance vis à vis other more complex methods. Our results also show the LDA rules perform surprisingly well even when the assumptions that underlie them do not hold.

Although our approach applies to general credit scoring situations—and to other screening and classification problems—it is particularly applicable to public utilities. (This work was part of a large study of credit granting to new telephone applicants by former AT&T operating telephone companies. The paper of Showers and Chakrin (1981) gives details of the entire project and its widespread implementation.) For concreteness consider a typical residential telephone applicant who calls the phone company business office to request telephone service. A service representative asks the customer a series of credit related background questions and based on the responses decides whether to request a security deposit from the customer. There are many thousands of such transactions each year. The customer's responses are recorded as a (vector) of numerical codes such as yes = 1, no = 0, or of actual values reported such as annual income = \$12,500, number of years at current residence = 6, etc. This response vector is then multiplied by a vector of weights to yield a credit score, and if the score is high enough the applicant "passes." In our application all response codes, as well as all weights were restricted to be zero or one to enable service representatives to compute a score by simply counting positive responses. Fortunately, this restriction turned out to have little effect on the power of the rules to discriminate between bad and good risks, and had other unanticipated benefits: The rules were very easy to explain and justify before public utility commissions, and it was relatively easy to include or delete a factor because of political or legal reasons.

A unique aspect of this study was that a very large sample of 87,000 customers was used for whom no prior screening was done. Thus we had a sample of *all* potential customers, not just those who had already been judged as good by some other method.

2. Decision Theory Framework

Credit screening can be viewed as a classical decision theory problem. In our application it suffices to hypothesize two states of nature representing the future payment behavior of a new customer: θ = (good, risk), and two actions that can be taken with each candidate customer: a = (pass, fail). The loss function $L(\theta, a)$ represents the consequence of taking action A when the true state of nature is θ and has the form

$$L(\text{good, pass}) = 0, \qquad L(\text{good, fail}) = \gamma \geqslant 0,$$

$$L(\text{risk, pass}) = \delta \geqslant 0, \qquad L(\text{risk, fail}) = 0. \tag{1}$$

A loss function with either L(good, pass) or L(risk, fail) nonzero can be reduced to an equivalent in the form of (1).

Information on the true state of nature is gained from observations on the random variables constituting the customer's *credit profile*, denoted by the vector $\mathbf{x} = (X_1, X_2, \ldots, X_J)$. Each X_j is the customer's response to a credit or background related question such as "Do you own a home?", "How long have you been employed?", "Do you have a credit card?", and so forth. The distribution $\mathbf{x}$ follows a probability function $f(\mathbf{x} \mid \theta)$ which depends on the true state of nature θ. If this distribution is substantially different for the good and the risk customers, profile data will provide useful information about the true state of nature θ. This is the key to the credit classification problem.[1]

[1] From this point on we treat the $\mathbf{x}$ vector as being specified and treat the problem of how best to use it. Clearly, the broader problem of what elements should belong in the $\mathbf{x}$ vector is also crucial. In practice, a variety of statistical methods are used heuristically to prune the dimension of the $\mathbf{x}$ to reasonable size.

A *decision rule* $d(\mathbf{x})$ determines the action to take when profile $\mathbf{x}$ is observed. Sometimes, as in our application, practical considerations restrict the form of the credit rule and we denote decision rules meeting these restrictions by the set Δ and consider only $d \in \Delta$. The performance of a rule d can be quantified by the *risk function* $R(\theta, d)$, the expected cost of using decision rule $d(\mathbf{x})$ when the true state of nature is θ. If we denote by $M_G(d)$ and $M_R(d)$ the probabilities of misclassifying a good and risk customer under rule d, respectively, then under the loss function (1), the risk function can be written as $R(\text{good}, d) = \gamma M_G(d)$ and $R(\text{risk}, d) = \delta M_R(d)$. Thus, with only two states of nature, the performance of a credit classification rule d can be represented as a point in the two-dimensional space $M_R(d)$ vs. $M_G(d)$ for which both axes range from 0 to 1. This change of scale finesses to some extent the fact that neither γ nor δ is likely to be known with great accuracy.

As discussed at length by Ferguson (1967) there is seldom a unique "best rule" in a statistical decision problem. Indeed we propose several formulations that suggest different ways of viewing the tradeoffs inherent in the credit granting situation.

Our first version postulates the additional knowledge of the "prior" probabilities π_g and $\pi_R = (1 - \pi_G)$ that a random customer is good or risk, respectively, and finds a Bayes rule minimizing expected losses:

P1: minimize$_{d \in \Delta} \gamma \pi_G M_G(d) + \delta \pi_R M_R(d)$.

Unfortunately, knowledge of both the loss function (γ and δ) and the priors (π_G and π_R) is typically difficult, if not impossible, to obtain. In public sector applications, the economic, political, or social consequences of the two different types of misclassification are quite different, even incommensurate. In private sector applications while it is at least in principle possible to represent both risks as dollar present values it is very difficult to accurately estimate the priors.

Our second formulation minimizes the probability of misclassifying a risk customer subject to a constraint on the probability of misclassifying a good customer:

P2: minimize$_{d \in \Delta} M_R(d)$ subject to $M_G(d) \leqslant \alpha$.

Here $\alpha \in [0, 1]$ is fixed, and we require knowledge of neither the priors nor the costs. Instead one must specify α, which may be imposed externally by, for example, the public utility commission. (Problem P2 is similar to the Neyman-Pearson approach to hypothesis testing, and it is the formulation we concentrate on later.)

A third formulation minimizes the probability of misclassifying a risk customer subject to a constraint on the overall probability of failing any customer:

P3: minimize$_{d \in \Delta} M_R(d)$ subject to $\pi_G M_G(d) + \pi_R[1 - M_R(d)] \leqslant \beta$.

Here $\beta \in [0, 1]$ is fixed.

A rule d that solves problem Pi is called efficient for Pi. Assuming π_g and π_R are fixed, we define the following three sets of efficient rules:

$E_1(\Delta) = \{d \in \Delta$: d solves P1 for some $\gamma/\delta \in [0, \infty]\}$,

$E_2(\Delta) = \{d \in \Delta$: d solves P2 for some $\alpha \in [0, 1]\}$,

$E_3(\Delta) = \{d \in \Delta$: d solves P3 for some $\beta \in [0, 1]\}$.

It is tempting to think that if a rule d solves P1 for some γ and δ, there is also an α for which it solves P2, and in addition a β for which it solves P3, and vice versa. However, such complete equivalence is not always true; the following theorems detail the valid relationships. (Proofs may be found in Kolesar and Showers 1983.)

THEOREM 1. $E_1(\Delta) \subseteq E_2(\Delta) \subseteq E_3(\Delta)$.

That is, every rule that is P1 efficient is also P2 efficient but the reverse inclusions do not always hold. To show when they do, we define the risk set $\Phi(\Delta) = \{\mathbf{y}: \mathbf{y} = [M_G(d), M_R(d)]$ for some rule $d \in \Delta\}$.

THEOREM 2. (a) *If* $\Phi(\Delta)$ *is convex, then* $E_2(\Delta) \subseteq E_1(\Delta)$.

(b) *If* $\Phi(\Delta)$ *is convex and* $(0, 1)$ (*pass everyone*) *is in* Δ, *then* $E_3(\Delta) \subseteq E_2(\Delta)$.

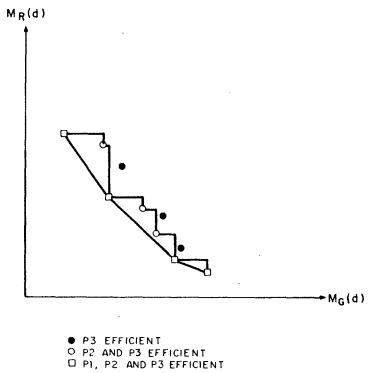

FIGURE 1. An Example of the Sets E1, E2 and E3.

When, as in our case, the data are binary Δ is discrete, and $\Phi(\Delta)$ is a nonconvex set. Thus, there are rules that are optimal for P2, but which are not optimal for P1 and rules which are optimal for P3 but not for P2. (See Kolesar and Showers 1983 for an example.)

Figure 1 gives examples in $M_R(d), M_G(d)$ space of all three types of efficient rules. The practical significance of the differences of course depends on "how far apart" these rules are and on how rich each family of rules is. We found that there were very few rules in E1, that E2 and E3 were similar and that E2 was easy to generate. Our desire to present decision makers with a rich set of rules to choose from reinforced our choice to solve P2 which was originally motivated by the ease of explaining the trade-off of failed "goods" versus passed "bads". If one allows randomized rules, the risk set D is convex, but unfortunately, particularly in "public" applications, randomized rules are inappropriate in credit screening since the same action is not always taken for customers with the same profile.

3. Review of the Standard LDA Approaches

Some advantages and disadvantages of the standard credit classification approaches led us to consider the problem afresh. (For additional background see Anderson 1958, Eisenbeis and Avery 1972, Goldstein and Dillon 1978, and Lachenbruch 1975. Cohen and Hammer 1966, Durand 1941 and Gentry 1974 give application of LDA to credit scoring.) The formulation usually solved is P1, and it is well known (Hoel and Peterson 1949) that P1 optimal rules are of the likelihood ratio type:

$$d(x) = \begin{cases} \text{"pass"} & \text{if } \dfrac{f(\mathbf{x} \mid \text{good})}{f(x \mid \text{risk})} \geqslant \dfrac{\delta \pi_R}{\gamma \pi_G}, \\ \text{"fail"} & \text{otherwise.} \end{cases}$$

Although this *looks* simple (pass customer profiles that have high ratios of good to risk customers), the actual structure of likelihood ratio rules is very complicated except for special densities $f(\mathbf{x} \mid \theta)$. Moreover, it can lead to the practical anomalies cited by Capon (1982). However, when $f(\mathbf{x} \mid \text{good})$ and $f(\mathbf{x} \mid \text{risk})$ are both multivariate normal densities, and have the same covariance matrix, the likelihood ratio reduces to the simple linear discriminant function. This is the basis for the optimality of the linear discriminant analysis (LDA) procedure since under these conditions and if the parameters are *known*, LDA is optimal for P1 (Wald 1944). But even when the data are multivariate normal with different covariance matrices, the optimal decision rule has a quadratic form. Moreover, in practice, the parameters are unknown, and it is never certain to which family of distributions $f(\mathbf{x} \mid \theta)$ belongs—indeed normality is rather doubtful. At *best*, one may have a sample of G known good customers and R known

risk customers, together with their profiles, $\mathbf{x}_l$, $l = 1, \ldots, G + R$. (At worst one only has data on customers that have earlier been screened by other procedures!) Nevertheless, most applications in the literature compute maximum likelihood estimates of the parameters which are then substituted for the true parameters in the linear discriminant function.

LDA is used even when the conditions for its optimality do not hold for several reasons: First, its theory is well developed and there exist readily available and very efficient LDA computer routines (BMDP 1977). Second, the resulting decision rule is relatively simple to implement. Third, the weights of the discriminant function depend only on the means and covariances, and thus, to take into account changes in cost estimates or priors, or simply to change the fraction of customers passing the rule, one need only adjust the cutoff. Fourth, LDA has produced rules that are better than purely subjective judgement and they have met a variety of legal challenges (Hsia 1978). And finally, since no one has reported an investigation of the sensitivity of the performance of LDA rules to aberrations in the underlying conditions, many practitioners proceed in the bliss of ignorance.

Since LDA is so often used when the normality and constant covariance conditions do not hold, we note that in these cases it actually maximizes the ratio of between to within group differences—assuming the within group differences are the same for both groups. This approach, which is Fisher's (1936) distribution-free approach to finding a linear discriminant function, is only obliquely related to P1, P2, or P3. It is, in short, a convenient and plausible surrogate problem.

Other standard classification techniques such as logit analysis, loglinear models, categorical partition analysis, and multinomial models which we found useful for heuristic selection of variables to include in $\mathbf{x}$ also exhibit similar disadvantages for actual rule construction: restrictive assumptions of distributional form, indirect objective function, or too complicated rule structure.

4. Mathematical Programming Approach

In the telephone company application, the conditions under which LDA solves P1 were clearly not satisfied. Most particularly, nearly all of the original data X_j, $j = 1, \ldots, J$ were binary. Another potential disadvantage of LDA was that the resulting weights $\mathbf{w}$ could have been any real number and anomalies such as those reported by Capon (1982) are not tolerable in a public forum. Finally, as mentioned earlier *very* simple rules were desired for operational reasons and we therefore restricted the weights to be 0 or 1.

To overcome the potential disadvantages of LDA cited above, we propose a mathematical programming approach. Since describing the background data ($\mathbf{x}$) with an explicit probabilistic model is difficult, if not impossible, we chose to work with the empirical sample distribution of the data, and treat it, in effect, as the true population distribution. This amounts to the (erroneous) assumption that one has sampled all customers to whom the rules will apply. However, with our very large sample of 87,000 customers this tactic seemed reasonable; indeed, such an empirical procedure is asymptotically optimal (Glick 1973).

Thus, we assume the availability of a large sample of K customers, each with known binary credit profile $\mathbf{x}$ of dimension J (there are J items on the questionnaire) and known actual credit behavior. There are $I = 2^J$ possible profiles. (In our application, ten yes-no questions were used so $I = 1024$.) Ideally, as was true for our sample, all prior screening rules or policies should be relaxed so that the sampled population includes a complete spectrum of potential customers. (An earlier stage in the analysis reduced the number of variables down from some 72 variables (Showers and Chakrin 1981).)

The data on the K customers can be summarized by the following quantities ($i = 1, \ldots, I$):

G_i = number of good customers with profile $\mathbf{x}_i$,

B_i = number of risk customers with profile $\mathbf{x}_i$.

We define the decision variables, p_i ($i = 1, \ldots, I$) as:

$$p_i = \begin{cases} 1 & \text{if profile } \mathbf{x}_i \text{ passes,} \\ 0 & \text{if profile } \mathbf{x}_i \text{ fails.} \end{cases}$$

Mathematical programming models corresponding to problems P1, P2, and P3 can be formulated directly. We call these formulations MP1, MP2 and MP3. We state only MP2, the model we worked with most:

$$\text{MP2: minimize} \quad \sum_{i=1}^{I} B_i p_i \qquad \text{subject to} \quad \sum_{i=1}^{I} G_i(1 - p_i) \leqslant \alpha \sum_{i=1}^{I} G_i,$$

where $p_i = 0$ or 1, $i = 1, \ldots, I$.

The values of p_i, $i = 1, \ldots, I$ that solve MP2 define the decision rule d^*:

$$d^*(x_i) = \begin{cases} \text{"pass"} & \text{if } p_i^* = 1, \\ \text{"fail"} & \text{if } p_i^* = 0. \end{cases}$$

Such formulations do not restrict the form of the rules; they simply partition the profile space into two subsets: those profiles that pass ($p_i = 1$) and those that fail ($p_i = 0$). No other partitioning can perform better on *this* sample of customers and therefore these models bound performance attainable on this sample of customers.

MP2 is a 0-1 knapsack problem which can be difficult to solve for large I. However, optimal solutions can be obtained very easily for *many* values of α by simply ordering the response vectors according to decreasing values of the ratio B_i/G_i. Imagine that this has been done and the profiles renumbered so that

$$B_i/G_i \geqslant B_{i+1}/G_{i+1} \qquad \text{for} \quad i = 1, 2, \ldots, M - 1,$$

where M is the number of profiles for which $B_i + G_i > 0$. A rule that requests deposits of persons with profiles with the (say) N highest values of this ratio will be exactly optimal for that α for which $\alpha \sum_{i=1}^{I} G_i$ equals $\sum_{j=1}^{N} G_j$. The "greedy" knapsack solutions obtained at these "exact points" also bound performance at intermediate points. Thus, if α_N and α_{N+1} are values at which the first N and $N + 1$ profiles are optimal, the performance for any $\alpha = \lambda\alpha_N + (1 - \lambda)\alpha_{N+1}$ with $0 < \lambda < 1$ is less than $\lambda \sum_{i=1}^{N} B_i + (1 - \lambda) \sum_{i=1}^{N+1} B_i$.

Note that ordering B_i/G_i and $B_i/(B_i + G_i)$ are the same and thus, for each value of α for which a "greedy" solution is exact, we have both P2 and P3 efficiency: For each $N = 1, \ldots, M$, there is a value of α and β for which the first N profiles are P2 and P3 optimal, respectively. Since the number of profiles with distinct B_i/G_i ratios in our sample was large, the greedy method produced a reasonably large set of optimal solutions over the many values of α and β, obviating the need for more refined algorithms to solve P2 and P3 at other α and β values.

The optimal knapsack solutions cannot however be implemented because even with large samples there will be many profiles with $B_i + G_i$ zero (i.e., no customers with that profile) or very small and one could not be *statistically* confident of the behavior of future customers with these profiles. Moreover, "knapsack" decision rules would be difficult to implement manually as they are essentially large tables indicating an action for each profile.

A modified mathematical programming approach that limits consideration to rules

with linear form overcomes the limitations of knapsack rules since such rules have a very simple structure *and* assume in effect that customers whose profiles are "close" to each other are likely to have similar credit behaviors. The following constraints force the rule to have a linear structure:

$$-Q(1-p_i) \leqslant \sum_{j=1}^{J} w_j x_{ij} - c \leqslant Qp_i - \epsilon, \qquad i = 1, \ldots, I.$$

Here Q and ϵ are constants and $\{w_j\}$ and c are variables. Q is a very large positive number. This approach has also been taken by Agin (1978). It can easily be checked that if

$$\sum_j w_i x_{ij} < c, \qquad p_i = 0, \quad \text{and if} \quad \sum_j w_j x_{ij} \geqslant c, \qquad p_i = 1.$$

Here the explicit decision variables are $w_j, j = 1, \ldots, J$ and c, from which the $\{p_i\}$ are determined implicitly.

Linear rules with continuous weights can be obtained with mixed integer programming codes such as MPSX (1979), but if the number of profiles with data is large (more than a few hundred), such an approach is very expensive. An algorithm tailored to take advantage of the problem's structure might be more efficient, but a reported attempt was not very successful (Agin 1978). When the weights must be zero or one the rules become a selection of a subset of N of the J profile elements and at least $n \leqslant N$ must appear in the profile of a passing customer. For reasonable J the number of possible such different rules is not *very* large, and the enumeration scheme given in the Appendix will solve the problem efficiently.

5. Nested Sets of Rules

We propose a consistency property for rules that has been important in public utility commission considerations; indeed its lack has been appropriately criticized by Capon (1982). Suppose a particular rule has been in use, and that a higher failure (deposit) level is now desired. It makes sense that persons failed under the old rule should also fail with a higher deposit level. Likewise, if the deposit level were reduced, customers who passed previously should still pass. We call this property *nesting*. Formally, let $f(d)$ be the fraction of customers failed under rule d, and let $F(d)$ and $F^c(d)$ be the set of profiles failed and passed under rule d, respectively. Let D be a set of rules. The set D is *nested* if for any d and d' in D, $F(d) \subseteq F(d')$ whenever $f(d) < f(d')$. We can show that LDA rules and the subset of knapsack rules generated by the greedy heuristic are nested. (In the following discussion we assume that all profiles are nonnegative vectors, that is, $\mathbf{x} \geqslant 0$ for all $\mathbf{x} \in X$. The following are stated without proof (see Kolesar and Showers 1983).)

LEMMA 1. *Let d and d' be linear rules with weights and cutoff $(\mathbf{w}, c)$ and $(\mathbf{w}', c')$ respectively. Suppose that $f(d) < f(d')$. Then d and d' are nested if $\mathbf{w} \geqslant \mathbf{w}'$ and $\mathbf{c} \leqslant \mathbf{c}'$.*

THEOREM 4. *The set of LDA rules is nested.*

THEOREM 5. *The subset of knapsack rules, generated by the greedy heuristic, is nested.*

Unfortunately, nesting is not a property that automatically holds for rules generated by mathematical programming methods. In particular, the set of all knapsack rules is not nested since, as the cutoff changes, individual profiles may be kicked out of the set of passing profiles and added back later. One might add constraints to the mathematical program to force the set of binary linear rules to be nested. We can show (Kolesar and Showers 1983) that:

THEOREM 6. *There can be no more than* $2J$ *distinct rules in a nested set of binary weighted linear rules.* (J *is the dimension of* **x**.)

Even without considering the computational burden of forcing mathematical programming rules to be nested, imposing this nesting constraint on a set of binary weighted linear rules is quite restrictive. With $J = 10$ factors, as was the case in the telephone company application, a nested set contains at most 20 rules whereas a nonnested set may contain hundreds. A relaxed concept of nesting may be appropriate in applications, and we suggest the following. It is common that customers fall into a small number of profiles. For these frequently occurring profiles (call this set L) we would like nesting to hold. In other words, if D is the set of rules and d and d' are in D and such that $f(d) < f(d')$, then we would like to have $F(d) \subseteq F(d') \subseteq L$. The fact that nesting does not hold for the "sparse" profiles not in L, for which there is little data upon which to base the action anyway, reflects our uncertainty about the correct action. In the telephone company study, binary weighted linear rules automatically satisfied this relaxed concept of nesting.

6. Evaluation Screening Methods

We compare several screening techniques using subsamples of the AT&T data and then make some general observations on implementation. In Sample A, profiles consisted of $J = 6$ binary questions and the sample contained 202 good and 154 risk customers. In Sample B, profiles consisted of $J = 7$ binary questions, and the sample contained 799 good and 527 risk customers. We compared (a) the Knapsack rules produced via the greedy heuristic, (b) the binary weighted linear rules produced via mathematical programming for problem P2 (the algorithm of §4 was used), and (c) rules produced via LDA. We computed the entire set of rules for all levels of failures for each technique and used performance plots of $1 - M_R$ vs. M_G to make our comparisons. The results for Sample A are shown in Figure 2a and those for Sample B are shown in Figure 3a. For clarity we have connected the points to produce the efficient curve for the particular method. (The line segments connecting efficient rules would represent the performance of rules which randomize between the extremes of the segments.) The steeper the curve, the better the rules. The dotted line running from (0, 0) to (1, 1) is for reference and represents the performance of "completely random" rules, which treat risk and good customers the same. Better rules lie above the dotted line.

In these figures, the "○" signs mark the performance of rules for values of α for which the greedy knapsack solution is exact and whose performance is a bound on the best that any type of rule can do on this data. Efficient curves for all other techniques

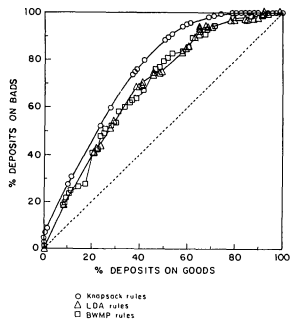

FIGURE 2a. Rule Performance on Development Sample A.

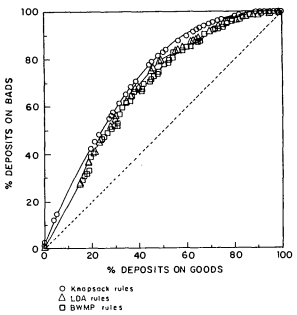

FIGURE 3a. Rule Performance on Development Sample B.

must lie between the knapsack curve and the dotted line. As mentioned before, a major problem in using these knapsack rules for screening procedures is insufficient data. This can be seen from viewing these figures. In Sample A only 43 out of a possible $2^6 = 64$ profiles were observed and in Sample B only 89 out of a possible $2^7 = 128$ profiles were observed.

To construct the linear discriminant analysis rules on these data the weights need be calculated only once. The entire set of rules is then constructed by varying the cutoff, c. The triangles in the figure represent cutoffs that caused a change in the performance.

The "□" marks the efficient curve for linear rules with binary (0, 1) weights constructed by the mathematical programming approach and the algorithm of §5. These rules perform comparably to the LDA rules; statistical tests would indicate no significant difference. It is important to note that by restricting the rules to be very simple (0 or 1 weights), the power of the rules to discriminate between the good and the risk customers has not been reduced. Comparing Figures 2a and 3a, we observe that the difference between LDA and binary weighted mathematical programming rules gets smaller and the difference between both of these techniques and the knapsack bound diminishes when the sample size is increased.

7. Comparison of Performance on a New Sample

No matter how well a rule performs on its "development sample," it is of little value unless it also performs well on *new* samples from the same population—actually on the future real world sample to which it would actually be applied. A "hold-out" sample can be used as a simulation of future performance in the nonstationary real world. To do this we used two additional samples drawn in an identical manner as the originals. The development samples (A and B) were used to construct the rules; the hold out samples (A′ and B) were used to determine if these results extrapolate reasonably. (Sample A′ contained 217 good and 154 risk, and Sample B′ contained 765 good and 533 risk.)

Figures 2b and 3b display the results. In these graphs "×"s mark the performance of the greedy knapsack rules constructed from the test sample data gives a bound on the best performance any technique can have on this sample. The performance of the techniques we are comparing must always be below this "×" curve. The other symbols refer to the rules constructed on the development samples "○" knapsack rules; "Δ" linear discriminant analysis rules; "□" binary weighted mathematical programming (BWMP) linear rules.

We first observe that although the knapsack rules are optimal for the sample used to construct them, they do poorly when extrapolated to new samples. This is because they

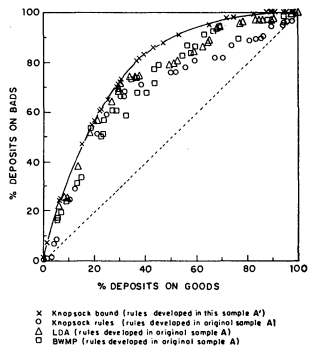

FIGURE 2b. Rule Performance on Test Sample A′.

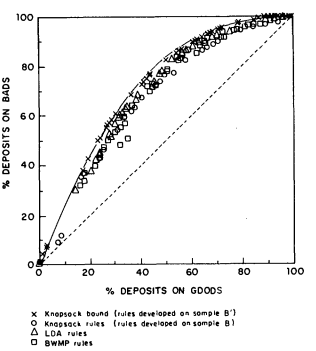

FIGURE 3b. Rule Performance on Test Sample B′.

request deposits of customers with profiles having monotonically decreasing values of R_i/G_i. The ordering of these ratios from the development sample is crucial, and in small samples this ordering will not be statistically stable. Thus, in Figure 2b, where the sample is small the knapsack rules do poorly in extrapolation, while with larger samples (Figures 3b) they do better in extrapolation. We also observe that linear discriminant analysis rules and binary weighted mathematical programming rules perform well on a new sample when the development sample is large (Figure 3b). However, when the development sample is small (Figure 2b), the performance of the binary weighted rules is erratic, since an unlikely random observation in a small development sample can significantly affect the rule developed. When the same observation does not occur in a new sample, the performance can be very different than predicted.

We conclude that "raw" knapsack rules are unsuitable for use on a new population because even with large samples there are problems of sparse data for many profiles. Of course if all profiles occurred sufficiently often, knapsack rules would perform well on a new sample and, since they are least restrictive of any procedure, would be optimal. The performance of LDA rules seems to be robust with respect to sample size and to the assumptions that assure their optimality. Knapsack and BWMP rules are more sensitive to random fluctuations in sample data than LDA rules. Note that because only estimates of means and covariances are needed to construct LDA rules, these rules are quite robust when extrapolated.

8. Summary and Conclusions

Unlike most of the literature on credit screening, this paper views the credit decision as a problem in multicriterion optimization. We have therefore proposed a multiplicity of appropriate problem formulations and solution methods. Our results show, at least in one particular large-scale implementation, that a variety of methods using different problem formulations and algorithms performs similarly. Our findings (1) support the use of linear discriminant analysis even when its assumptions appear to be violated, and (2) show that simpler binary methods are also quite justified and can have distinct advantages. The latter finding is of practical importance because rules with simple structure are very easy to implement on a mass scale and have enough appeal to survive the sometimes arduous public approval processes. Our "n out of N" rules have proven to be both effective screens and robust to "external" social or political value judgments. For example, in one jurisdiction although "car ownership" did not fall in the class of optimal rules, a public interest group lobbied for it. It was easy to include and doing so did not measurably diminish the quality of the screening. In another jurisdiction the externally imposed requirement that senior citizens get a positive weight was accommodated—also without appreciable negative impact.

The sorting algorithm proposed here generates nearly optimal linear screening functions with binary weights. As a result, only a simple count of the number of positive responses to the credit questionnaire is needed to reach a credit decision. Moreover, rather than producing a single "optimal" solution the algorithm easily produces an efficient family of "n out of N" credit screens. In the actual implementation the differences between the utilities', consumer groups' and public service commission's objectives have been reasonably adjudicated by selection from among this family of efficient solutions—sometimes with minor adjustments.

Appendix

The restriction to binary weights greatly simplifies implementation. Simply count the number of positive responses. With binary weights $\{w_j\}$, $j = 1, \ldots, J$, there are 2^J possible rules and, if the number of questions, J, is not too large, the following enumeration scheme will efficiently identify *all* P2 optimal rules.

A slight modification will produce P3 optimal rules. P1 rules are difficult to generate directly, but since P1 rules are P2 optimal they could be identified subsequently.

We define the several counts of customers in the sample: $rp(d)$ is the number of risk customers that pass rule d and $gf(d)$ is the number of good customers that fail rule d; R, and G are the total number of risk customers and good customers, respectively. Decision theoretic and relationships that underlie the algorithm are developed in Kolesar and Showers (1983).

The algorithm is:

Step 1. Initialize $A = \{d_0, d_\infty\}$. d_0 is the rule that passes all customers. It has all 0 weights and cutoff 0. ($rp(d_0)$ = total number of risk customers, $gf(d_0) = 0$.) d_∞ is the rule that fails all customers. It has all unit weights and cutoff ∞. ($rp(d_\infty) = 0$, $gf(d_\infty)$ = total number of good customers.)

Step 2. Let d be a new rule. (If all rules have been considered, then stop.) If there exists an i such that $gf(d_{i-1}) < gf(d) < gf(d_i)$, then go to Step 3. Otherwise, there exists an i such that $gf(d_i) = gf(d)$. Go to Step 4.

Step 3. Let i be such that $gf(d_{i-1}) < gf(d) < gf(d_i)$. If $rp(d_{i-1}) \leqslant rp(d)$, throw out d and go to Step 2. If $rp(d) \leqslant rp(d_i)$, find the largest j for which $rp(d) \leqslant rp(d_j)$ and throw out d_i through d_j and their equivalent sets. Insert d in the set A just after d_{i-1}, call it rule d_i, and re-index rules $j+1, j+2$, etc., as rules $i+1$, $i+2, \ldots$. Go to Step 2.

Step 4. Let i be such that $gf(d_i) = gf(d)$. If $rp(d_i) < rp(d)$, throw out d and go to Step 2. If $rp(d_i) = rp(d)$, add d to d_i's equivalent set $E(d_i)$ and go to Step 2. If $rp(d) < rp(d_i)$, find the largest j for which $rp(d) < rp(d_j)$, throw out d_i through d_j and their equivalent sets, replace d_i in the set A with d, re-index rules d_{j+1}, d_{j+2}, etc., as rules d_{i+1}, d_{i+2}, etc.

Note that implementation of the algorithm is quite straightforward with a list processing routine which automatically takes care of the bookkeeping involved in insertion and deletion of rules from the "list" A.

References

AGIN, N., "An Optimization Approach to Credit Scoring," presented at the joint TIMS/ORSA Meeting, May 1978 (mimeo), Mathtech, Princeton, N.J., 1978.

ANDERSON, T. W., *An Introduction to Multivariate Statistical Analysis*, John Wiley and Sons, Inc., New York, 1958.

Biomedical Computer Programs (BMDP-77), W. J. Dixon (Ed.), University of California Press, Berkeley, 1977.

CAPON, N., "Credit Scoring Systems: A Critical Analysis," *J. Marketing*, 46 (1982), 82–91.

COHEN, K. J. AND F. S. HAMMER, *Analytical Methods in Banking*, Chapter 6, Richard D. Irwin, Inc., Homewood, Ill., 1966.

DURAND, D., "Risk Elements in Consumer Installment Financing," *Studies in Consumer Installment Financing: Study 8*, National Bureau of Economic Research, 1941.

EISENBEIS, R. A. AND R. B. AVERY, *Discriminant Analysis and Classification Procedures: Theory and Applications*, Lexington Books, Lexington, Mass., 1972.

FERGUSON, T. S., *Mathematical Statistics: A Decision Theoretic Approach*, Academic Press, New York, 1967.

FISHER, R. A., "The Use of Multiple Measurement in Taxonomic Problems," *Ann. Eugenics*, 7 (1936), 179–188.

GENTRY, J. W., "Discriminant Analysis in the Credit Extension Decision," *Credit World*, 63 (1974), 25–28.

GLICK, N., "Sample Based Multinomial Classification," *Biometrics*, 29 (1973), 241–256.

GOLDSTEIN, M. AND W. R. DILLION, *Discrete Discriminant Analysis*, John Wiley and Sons, New York, 1978.

HOEL, P. G. AND R. P. PETERSON, "A Solution to the Problem of Optimum Classification," *Ann. Math. Statist.*, 20 (1949), 433–438.

HSIA, D., "Credit Scoring and the Equal Credit Opportunity Act," *Hastings Law J.*, 30 (1978), 371–448.

IBM Mathematical Programming System Extended/370 (MPSX/370), General Information Manual, GH19-1090-3, IBM, 1979.

KOLESAR, P. AND J. L. SHOWERS, "Credit Screening with Categorical Data: A Decision Theoretic-Mathematical Programming Approach," Columbia Business School Working Paper, No. 393, 1983.

LACHENBRUCH, P. A., *Discriminant Analysis*, Hafner Press, New York, 1975.

LUENBERGER, D. G., *Introduction to Linear and Nonlinear Programming*, Addison-Wesley, Reading, Mass., 1973.

MORRISON, D. F., *Multivariate Statistical Methods*, 2nd Ed., McGraw-Hill, New York, 1976.

MYERS, J. H. AND E. W. FORGY, "The Development of Numerical Credit Evaluation Systems," *J. Amer. Statist. Assoc.*, 58 (1963), 799–806.

ORGLER, Y. E., *Analytical Methods in Loan Evaluation*, Lexington Books, Lexington, Mass., 1975.

SHOWERS, J. L. AND L. CHAKRIN, "Reducing Uncollectible Revenue from Residential Telephone Customers," *Interfaces*, II-6 (1981), 21–34.

WALD, A., "On a Statistical Problem Arising in the Classification of an Individual into One or Two Groups," *Ann. Math. Statist.*, 15 (1944), 145–162.