better classifier the scorecard is. Define the Gini coefficient G to be twice this area. This has the useful property that the perfect classifier that goes through A in Figure 7.3 will have $G = 1$, while the random classifier with ROC curve OB in Figure 7.3 has $G = 0$. In fact, one can show that it is the probability that a randomly selected bad will have a lower score than a randomly selected good. This is related to the Wilcoxon test for two independent samples (Hand 1997). Thus the Gini coefficient gives one number that summarizes the performance of the scorecard over all cutoff scores. This is both useful in its brevity and misleading because in reality we are usually interested in the scorecard's performance over a small range of possible cutoff scores. The same complaint can be made of the Kolmogorov–Smirnov statistic and the Mahalanobis distance. They all describe general properties of the scorecard, whereas what is important in practice is how the scorecard performs at the chosen cutoff.

One can use the ROC curve to identify suitable cutoff scores. The score that maximizes the Kolmogorov–Smirnov statistic, for example, corresponds to the point on the curve whose horizontal distance from the axis is greatest. This is C in Figure 7.4. This follows since the point is $(P_B(s), P_G(s))$, so this horizontal distance is $(P_B(s) - P_G(s))$. If one assumes that the true percentages of goods and bads in the population are p_G and p_B, respectively, and that L and D are the usual loss quantities for misclassifying, the expected loss rate if one has a cutoff at s is

$$l(\text{Actual}) = LP_G(s)p_G + D(1 - P_B(s))p_B. \tag{7.17}$$

As far as an ROC curve f is concerned, this is like having to minimize $Lp_G f(x) + Dp_B(1-x)$, which occurs when

$$Lp_G f'(x) - Dp_B = 0 \tag{7.18}$$

(i.e., the derivative is 0). Hence at the minimum the slope of the tangent to the ROC curve, $f'(x)$, satisfies $f'(x) \quad \frac{Dp_B}{Lp_G}$. One way to find this point is to draw the line with slope $-\frac{Lp_G}{Dp_B}$ through the point $(1, 0)$ and project the curve onto this line. The point that yields the projection nearest the point $(1, 0)$ is the point we require. This is point D in Figure 7.4.

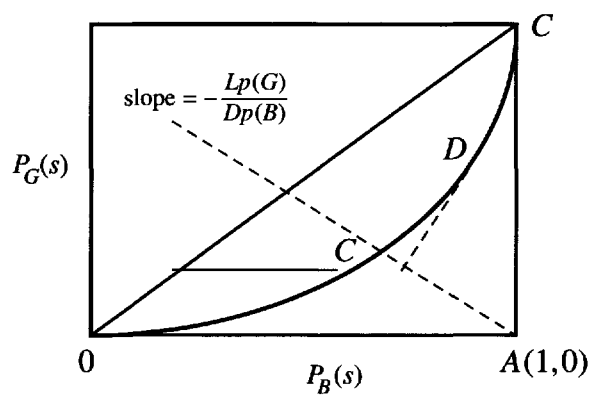


Figure 7.4. *Using ROC curves to find cutoffs.*

7.7 Comparing actual and predicted performance of scorecards: The delta approach

The previous sections of this chapter looked at ways to measure the classification performance of a scoring system and the overall characteristics of a scorecard, but they did not seek to identify where and why it was misclassifying. Those measures are used at the development of the scoring system to describe its likely performance characteristics and are used to change cutoff levels. In this section, we look at one technique that identifies the differences between the way the scorecard is performing in practice and how it was intended to perform. The different forms of monitoring reports, which also perform this function, are discussed in Chapter 9. The delta approach has the advantage of calculating a measure for the misalignment between the predicted and actual scorecard performance and splitting that to identify the misalignment in different scoring reasons. It can then be used to decide whether the scorecard is satisfactory or needs to be completely rebuilt or whether adjusting one or two scores will be sufficient. It can be used only on scoring systems based on regression methods, and we assume hereafter that the original scorecard was built using logistic regression.

The original scorecard was built assuming a model that

$$\begin{aligned} &P(G|\text{Score} = s) = F(s), \\ &\quad \text{where Score} = w_0 + w_1x_1 + \cdots + w_px_p \quad \text{and} \quad F \text{ is a monotone function,} \end{aligned} \tag{7.19}$$

and traditionally one wished to test whether $X_i = 1$, say, had the correct score by looking at

$$P(G|s_1 < \text{Score} \leq s_2) = F(s_2) - F(s_1) = P(G|s_1 < \text{Score} \leq s_2 \text{ and } X_i = 1). \tag{7.20}$$

Doing this directly can be very time-consuming since there are so many choices of s_1, s_2, i, and values for X_i. An alternative approach is to exploit knowledge of $F(s)$. In logistic regression,

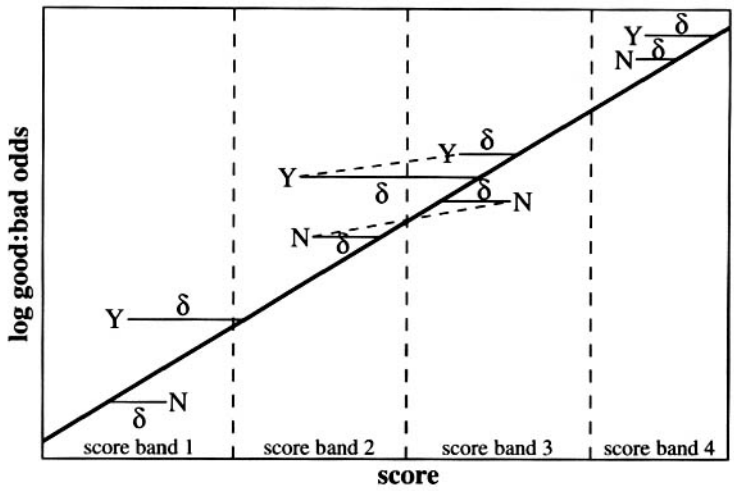
$$F(s) = \frac{e^{\alpha+\beta s}}{1 + e^{\alpha+\beta s}},$$

so

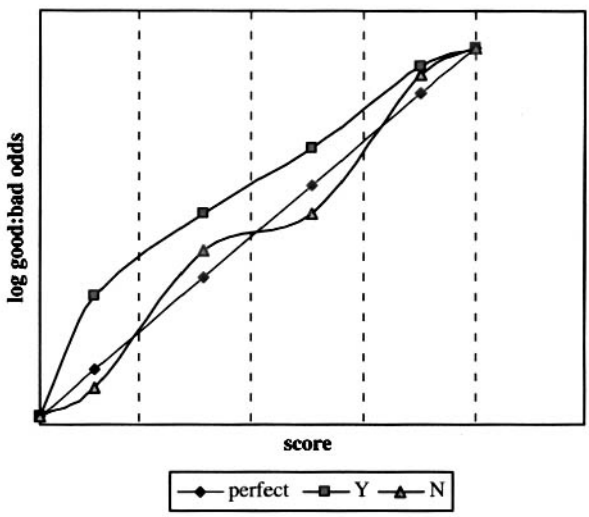
$$\log\left(\frac{P(G|\text{Score} = s)}{P(B|\text{Score} = s)}\right) = \log(\text{Odds}(G : B)|s) = \alpha + \beta s. \tag{7.21}$$

What we do is plot the log of the actual good:bad odds in different score ranges for different attributes of the characteristic under consideration. Suppose that one is considering telephone ownership, with Y meaning the consumer has a telephone number and N meaning the consumer does not have a number. Figure 7.5 shows the graph of these odds when the score range has been split into four ranges, and so each graph consists of joining the four points of mean score and log of actual odds in each score range. If the scorecard were perfect, then each of the graphs would lie on the line $\alpha + \beta s$. That is not the case—the telephone owners have better odds than predicted and the nonowners have worse odds in three of the four score ranges. The delta score at each point is the amount needed to be added to the score so that the point coincides with the original graph:

$$\delta = \frac{\log(\text{Actual odds}) - \log(\text{Predicted odds})}{\beta} = \frac{\log\left(\frac{\text{Actual odds}}{\text{Predicted odds}}\right)}{\beta}. \tag{7.22}$$



(a)



(b)

Figure 7.5. (a) *Graph of odds against score and corresponding delta distances.* (b) *Log odds against score in four score bands.*

Assume that one has a set of data of the actual performance of the scorecard, i.e., which customers are good and bad and what their original score (and the β_0 in the original logistic regression) was. Let g_{ij} (b_{ij}) be the actual number of goods (bads) in the ith score band with attribute j for a given characteristic and g_i (b_i) be the total number of goods (bads) in the ith score bands. Apply another logistic regression to this data to estimate for each original score the predicted probability of a customer with that score being bad, using the actual data; i.e., one is fitting

$$P(B|\text{Score} = s) = \left(\frac{1}{1 + \exp\left(\hat{\alpha} + \hat{\beta}s\right)}\right). \tag{7.23}$$

Let $\hat{b}_{ij}, \hat{b}_i$ be the sum of the predicted badness for those in the ith score band with attribute j and those in the ith score band in total.

Estimate the actual good:bad ratios by

$$r_{ij} = \frac{g_{ij} + \frac{1}{2}}{b_{ij} + \frac{1}{2}}, \qquad r_i = \frac{g_i + \frac{1}{2}}{b_i + \frac{1}{2}}. \tag{7.24}$$

(These are better estimates of $\log r_{ij}$ than taking $r_{ij} = \frac{g_{ij}}{b_{ij}}$.)

Estimate the predicted good:bad ratios by

$$\hat{r}_{ij} = \frac{g_{ij} + b_{ij} - \hat{b}_{ij}}{\hat{b}_{ij}} \quad \text{and} \quad \hat{r}_i = \frac{g_i + b_i - \hat{b}_i}{\hat{b}_i}. \tag{7.25}$$

It is possible that $\hat{r}_i$ and r_i differ considerably although all $\hat{r}_{ij}$ are consistent with $\hat{r}_i$. This means the whole curve has changed somewhat although the scores for the attributes under consideration are consistent. We allow for this by defining a scaled predictor good:bad ratio to be

$$\bar{r}_{ij} = \frac{\hat{r}_{ij} r_i}{\hat{r}_i}. \tag{7.26}$$

This adjustment ensures that the centroid of the predictor line in the ith band agrees with the centroid of the line obtained in the original logistic regression.

Then define δ:

$$\delta_{ij} = \frac{\log\left(\frac{r_{ij}}{\bar{r}_{ij}}\right)}{\beta}.$$

One can show that δ_{ij} has an approximate standard error of

$$\sigma_{ij} = \frac{\sqrt{\frac{g_i - g_{ij}}{g_i\left(g_{ij} + \frac{1}{2}\right)} + \frac{b_i - b_{ij}}{b_i\left(b_{ij} + \frac{1}{2}\right)}}}{\beta}, \tag{7.27}$$

and if one takes the weighted mean of the delta scores in region i,

$$\bar{\delta}_i = \frac{\frac{\sum_j \delta_{ij}}{\sigma_{ij}^2}}{\sum_j \frac{1}{\sigma_{ij}^2}},$$

one can apply a χ^2 test,

$$e^i = \sum_j \frac{(\delta_{ij} - \bar{\delta}_i)^2}{\sigma_{ij}^2}, \tag{7.28}$$

to check if the scorecard is badly aligned in the ith score band. Summing the e^i's gives a measure of the overall alignment of the scorecard.

If one wanted to adjust the scores to improve the alignment, one would add the aggregated delta scores for attribute j, δ_j to the current score for j, where

$$\delta_j = \frac{\sum_j \frac{\delta_{ij}}{\sigma_{ij}}}{\sum_j \frac{1}{\sigma_{ij}}}. \tag{7.29}$$

This would be a way to get an overall adjustment to that attribute, but the δ_{ij} give a more detailed insight into what are the errors caused by that attribute at different ranges of the scorecard.

Chapter 8

Practical Issues of Scorecard Development

8.1 Introduction

The previous chapters looked at the different techniques that can be used to develop application and behavioral scoring systems and ways to measure the performance of such systems. This chapter is somewhat more down to earth and concentrates on the practicalities of building a scoring system using these techniques, and it identifies some of the pitfalls that can arise.

Section 8.2 considers how to select the sample of previous customers on which to build the system. Section 8.3 discusses the usual definitions of good and bad performance for such borrowers and what one does with those in the sample who do not fall into either category. Sections 8.4 and 8.5 consider the characteristics that tend to be used in building scoring systems, the first section concentrating on what application and transactional data are used and the second on the credit bureau information available. The next three sections concentrate on how these characteristics are modified and then used in scoring systems. Section 8.6 discusses how and why one might develop a suite of scorecards—one for each of the subpopulations into which the sample is divided. Section 8.7 looks at coarse classifying the characteristics—putting several attributes into the same group so they get the same score—and splitting the values of a continuous variable into a number of different categories. Section 8.8 discusses how one chooses which variables to keep in a scorecard.

Section 8.9 addresses the thorny issue of reject inference; i.e., How can one enhance the scoring system by using the partial information available on those previous applicants who were rejected for credit? Section 8.10 discusses how and why the decisions of the scoring system may be overridden and how that affects the performance of the scoring system, while Section 8.11 concentrates on how the cutoff score in the scorecard is chosen. Section 8.12 investigates how one checks if the scorecards are correctly aligned and how to recalibrate them if necessary.

8.2 Selecting the sample

All the methodologies for credit and behavioral scoring require a sample of previous customers and their histories to develop the scoring system. There are two somewhat conflicting objectives in selecting such a sample. First, it should be representative of those people who are likely to apply for credit in the future—the through-the-door population. Second, it

should incorporate sufficient of different types of repayment behavior (i.e., the goods and the bads) to make it possible to identify which characteristics reflect this behavior in the general through-the-door population. The group that must be closest to this population are the previous applicants for that lending product. The conflict arises because to get as close as possible to the future through-the-door population, we want the sample group to be as recent as possible. However, to distinguish between the good and bad repayment behavior, we need a reasonable history of repayment and thus a reasonable time since the sample group applied. This is particularly the case with behavioral scoring, where one also needs a reasonable performance period to identify the transactional characteristics as well as an outcome period. The compromise is usually an outcome period of 12 months for application-scoring systems. In behavioral scoring, one usually takes 18 to 24 months of history and splits that into 9 to 12 months of performance history and 9 to 12 months of an outcome period. These periods vary depending on the product since for mortgages the outcome period may need to be several years.

The next questions are how large the sample should be and what the split should be between the number of goods and the number of bads in the sample. Should there be equal numbers of goods and bads in the sample, or should the sample reflect the good:bad odds in the population as a whole? Normally, the latter is so strongly oriented to the goods (say, 20:1) that keeping the same odds in the sample would mean there may not be enough of a bad subpopulation to identify their characteristics. For this reason, the sample tends to be either 50:50 or somewhere between 50:50 and the true population proportion of goods to bads. If the distribution of goods and bads in the sample does not have the same distribution as that in the population as a whole, then one needs to adjust the results obtained from the sample to allow for this. In the regression approach, this is done automatically as the probability of goods and bads in the true population, p_G, p_B, is used in the calculations. In other approaches, it has to be done a posteriori, so if a classification tree built on a sample where the goods were 50% of the population (but from a true population they were 90%) has a node where the good bad ratio is 3:1 or 75% to 25%, the true odds are

$$\frac{(\text{odds in node}) \cdot (\text{odds in true population})}{(\text{odds in sample population})} = \frac{\frac{3}{1} \cdot \frac{9}{1}}{\frac{1}{1}} = 27{:}1.$$

As for the number in the sample, Lewis (1992) suggested that 1,500 goods and 1,500 bads may be enough. In practice, much larger samples are used, although Makuch (1999) makes the point that once one has 100,000 goods, there is no need for much more information on the goods. Thus a typical situation would be to take all the bads one can get into the sample and take 100,000+ goods. This sample is then randomly split into two. One part is used for developing the scoring system and the other is used as a holdout sample to test it.

The real difficulty arises when samples of such sizes cannot be obtained. Is it sensible to put applicants for different products or applicants to different lenders together to make an enlarged sample? This is discussed in more detail in section 12.2, where we look at the development of generic scorecards, but it is sufficient to say that one has to be very careful. For example, the attributes that distinguish risky borrowers for secured loans like mortgages or even hire purchase of white goods are different from those that identify the risky borrowers for unsecured loans.

If the sample is chosen randomly from an existing population of applicants, one has to be sure the choice is really random. This is not too difficult if there is a central application list kept with applicants ordered by time of application. Choosing every tenth good in the list should give a reasonably random sample of 10% of the goods. If, however, one has to

go to branch levels to collect the list, one has to first randomly select the branches to ensure a good mix of urban and rural branches and a suitable spread of socioeconomic conditions and geography. Then one has to randomly select at the branch level. Again one needs to be careful. Deciding to take all the customers who apply in a given month might on the surface seem sensible, but if that is the month when universities start back, then a lot more students are likely to be in the applicant population than in a true random sample. Sometimes it may be necessary to put such a bias into the sample since, for example, the new product is more for young people than the existing one and so one wants a higher proportion of young people in the sample than in the original population. The aim always is to get a sample that will best reflect the likely through-the-door population for the new product. In fact, this is not quite correct; what one wants is a sample of the through-the-door population that the lender will consider for the scoring process. Thus those who would not be given loans for policy reasons should be removed from the sample, as should those who would be given them automatically. The former might include underage applicants, bankrupts, and those with no credit bureau file. The latter might include customers with specific savings products or employees of the lender.

All this work on the choice of sample presupposes that we can define goods and bads. In the next section, we consider what these definitions might be and what to do with those in the sample who do not obviously fall into one or the other category.

8.3 Definitions of good and bad

As part of the development of a scorecard, one needs to decide how to define good and bad. Defining a bad does not necessarily mean that all other cases are good. Often in scorecard development, at least two other types of case can be identified. The first might be labeled "indeterminates," those cases that are in between—neither good nor bad. The second might be labeled "insufficient experience."

In a scorecard development for a credit card portfolio, a common definition of bad is a case that at some point becomes three payments in arrears. This is often referred to as "ever 3+ down" or "worst 3+ down." The indeterminate cases might be those that have a worst-ever status of two payments down. Thus they have caused some problems and some additional collections activity—perhaps repeatedly if they have been two payments down on several occasions—but have never become three payments down. Then we might identify those where we have insufficient experience. Suppose that we have a development sample window of 12 months of applications and an observation point one year later, so cases have 12 to 24 months' exposure. Then we might label as having insufficient experience those with three or fewer months of sales or cash advance activity. In other words, the account is not bad, but it has been used fairly infrequently and so it would be premature to adjudge it to be good. The residue would be categorized as good accounts.

This classification is only an example. Many variations are possible. One could define bad as being ever 3+ down or twice 2 down. We could define "insufficient experience" as those that have had less than six months of debit activity. We could include in the indeterminate grouping those cases that have missed one payment.

On an installment loan portfolio, the situation might be a little clearer. Here, we might define bad as ever 3+ down or ever 2+ down or something in between, as suggested above. Indeterminates would be those that may have missed one payment, or may have missed one payment several times, or may have missed two payments. If we have chosen a sample window carefully, then there may be no cases with insufficient experience. However, if

someone redeems a loan, i.e., repays a loan in a lump sum, after only a few months, we could classify the loan into this category, especially if the truncated loan means that we did not make a profit or made a very small profit. The survival analysis idea, introduced in section 12.7, is one approach to dealing with the difficulty that different outcomes are unsatisfactory as far as the profitability of a loan is concerned.

When we move into secured lending, such as a mortgage, our definitions could change markedly. Here we have some security and so our definition of good and bad may be affected dramatically by whether the case generated a loss. Thus if a mortgage defaults and the property is repossessed and we recover all our loan and our collection and litigation costs, one could argue that this is not a bad account. Some lenders would classify this as bad in the scorecard development, while others may consider indeterminate to be a more appropriate label. Similar to the installment loan, if a mortgage is redeemed fairly early in its life, we may wish to classify this as indeterminate or insufficient experience.

In the case of current (checking) accounts with an overdraft, by necessity the definitions again have to be changed. There is no fixed expected monthly repayment, so an account holder cannot be in arrears. Therefore, a completely different set of definitions needs to be introduced. We might categorize a bad account as one where there is unauthorized borrowing, i.e., borrowing above the agreed overdraft limit, if there is one.

Whatever definitions we choose, there is no effect on the scorecard methodology. (This assumes that the definitions create a partition; i.e., all possible cases fall into exactly one classification.) We would normally discard the indeterminates and those with insufficient experience and build the scorecard with the goods and bads. Of course, how we define goods and bads will clearly have an effect on the result of the scorecard development. Different definitions may create different scorecards. These differences could be in the score weights or in the actual characteristics that are in the scorecard. However, that does not mean that the results will be very different. Indeed, different definitions of good and bad may generate quite different scorecards but still result in a great similarity in the cases that are accepted and declined.

Another issue that might arise is if we use a very extreme definition of bad. Here we may end up with few actual bad cases and so our modeling robustness may suffer. At a practical level, the definition of good and bad should be considered as more than a theoretical exercise. Issues of profit or loss may have a bearing. We also want consistency in the definition of good and do not want to use a factor to help classify accounts that is highly variable. In general, however, while it is important to develop sound and reasonable definitions of good and bad, it may make only a marginal difference to the effectiveness of the scorecard actually developed based on the definitions.

8.4 Characteristics available

The characteristics available to discriminate between the good and the bad are of three types—those derived from the application form, those available from a credit bureau search, and, for behavioral scoring only, those describing the transactional history of the borrower. We deal with the credit bureau data in the next section, so we concentrate here on the application characteristics

Table 8.1 shows the characteristics that occur in three typical application forms: one from a finance house for a car loan, a U.S. credit card, and a U.K. credit card.

Several characteristics are not permitted to be used for legal reasons. The U.S. Equal Credit Opportunity Acts of 1975 and 1976 made it illegal to discriminate in the granting

Table 8.1. *Characteristics in three application forms.*

Characteristic	Finance house	U.S. credit card	U.K. credit card
ZIP/postal code	X	X	X
Time at address	X	X	X
Residential status	X	X	X
Occupation	X	X	X
Time at employment	X	X	X
Applicant's monthly salary	X	X	X
Other income	X	X	
Number of dependents	X	X	
Number children	X	X	X
Checking account/current account	X	X	X
Savings account	X	X	
Credit card	X	X	X
Store card	X	X	X
Date of birth			X
Telephone		X	X
Monthly payments	X		
Total assets	X		
Age of car	X		

of credit on the grounds of race, color, religion, national origin, sex, marital status, or age. It is interesting to note, however, that one can use age in U.S. score cards if the scores given to ages of 62+ are greater than those given to any other age. The U.K. Race and Sex Discrimination Act outlaws discrimination in credit granting on the grounds of race or sex. Other variables like health status and previous driving convictions, although not illegal, are not used because lenders consider them culturally indefensible.

Having decided on the application questions or characteristics, one also has to decide on the allowable answers or the attributes. For example, in terms of residential status, should one leave the answer open where one could have answers like "owner with no mortgage," "with friends," "barracks," "caravan," etc., or should one restrict the answers to choosing one from "owner/renting/with parents/others"? Another question is how one decides what to do when no answer is given. The usual approach is to have a category of "answer missing" for each characteristic, but there are occasions where it may be clear that no answer really means "no" or "none" (e.g., in response to the spouse's income question).

The occupation characteristics can cause real problems since it is very hard to code because there are hundreds of thousands of occupations. One approach is to specify the attributes such as executive, manual worker, manager, etc. Even with this, "manager" can cause no end of problems as one may be managing a factory employing thousands or managing a career as a part-time Santa Claus. If one goes to the other extreme and simply asks if the applicant is employed, self-employed, or unemployed, then many people can answer yes to at least two of these answers.

Income can also cause difficulty unless one is very careful in the wording of the question to clarify that it refers to income per month (or per year) and whether it is the applicant's basic income, the applicant's total income, or the total household income.

Finally, it is necessary to be able to validate the data. One needs to check that there are no ages under 18 and not too many over 100. For the different characteristics, one can analyze the distribution of the answers to check that they seem to make sense. The verification of the data is also a way of trying to identify fraud—it is amazing how often inconsistent or impossible answers are given. One of the other problems that can arise is when an applicant

fills in the application form with help from the lender's staff. If the staff have incentives for signing up customers that are accepted for loans, they might be tempted to advise the applicant on suitable answers. This is why some lenders use the credit bureau data as their main source of application information, although such lenders are in a minority.

For behavioral scoring, one can add a large number of transactional characteristics. The most common would be average balance and maximum and minimum balance over the past month, the past 6 months, or the past 12 months. Other characteristics could include the total value of credit transactions and the total value of debit transactions over such periods. One would also need to include characteristics that suggest unsatisfactory behavior, like the number of times over the credit card or overdraft limit or the number of reminder letters that have had to be sent. One can produce a large number of variables from transactional data, and these are likely to be strongly correlated with each other. Deciding which to keep and which to ignore is part of the art of scoring.

8.5 Credit bureau characteristics

Credit reference agencies or credit bureaus exist in many countries. Their roles are not identical from country to country, and neither are the legislative frameworks in which they operate. Therefore, it should come as no surprise that the stored and available data vary from country to country and even within countries.

In section 2.10, we introduced, at a fairly high level, the role of the credit reference agency or credit bureau. In this section, we shall describe in some detail what is available from credit bureaus in the U.K. This is to give the reader some appreciation of the extent of the information that is collected, analyzed, and made available. Many other Western countries operate in a similar way.

In the U.K., there are two main bureaus for consumer information—Experian and Equifax. Information is accessed through name and address, although there are different levels of these being matched (see later). The information they have available on consumers falls into several types, and we deal with each one in turn:

- publicly available information,
- previous searches,
- shared contributed information—through closed user groups, many lenders share information on the performance of their customers,
- aggregated information—based on their collected information, such as data at postcode level,
- fraud warnings,
- bureau-added value.

8.5.1 Publicly available information

In the U.K., publicly available information is of two types. The first is the electoral roll or voters' roll. This is a list of all residents who have registered to vote. In the U.K., this is not a complete list of all adults since there is no requirement to vote, as there is in some other countries. This information also includes the year in which someone was registered to vote at an address. This is useful information because it can be used to validate the time that

someone has stated that they've lived at an address. For example, if they state that they have lived at an address for three years but have been registered to vote there for twelve years, there is clearly a mismatch of information.

As this is being written, there is a debate in the U.K. between the Office of the Data Protection Registrar (ODPR) and the regulatory and industry bodies. This debate has arisen because the ODPR wishes voters to be able to decide whether their voter registration should be available to be used for credit or marketing purposes. It looks likely that some constraints will be implemented. However, while this is being done to provide greater protection for the consumer, it is possible that the effect will be to weaken the lenders' ability to distinguish between acceptable and unacceptable credit risks. If so, the consumer may actually suffer either by being declined or by being forced to pay more in interest rates to cover the additional risks.

There is no legal obligation on local councils to provide the electoral roll to the bureau. The electoral roll is available on paper to political agents with regard to an election and is often available for inspection in local libraries or council offices. However, in almost all cases, the electoral roll is supplied to the bureaus electronically in return for a considerable fee.

The second type of public information is public court information. These are the details of county court judgements (CCJs) or, in Scotland, court decrees. One option as part of the process to pursue a debt is to go to a county court and raise an action for a CCJ. This establishes the debt and may force the debtor into action. If the debtor then clears the debt and this information is passed to the bureau, the CCJ does not disappear but will show as having been satisfied. Similarly, if there is a dispute concerning the CCJ, which is proved in favor of the plaintiff, the CCJ may show as being corrected. This may occur, for example, where the court action was raised for a joint debt when the liabilities of one of the debtors had been extinguished. In such a case, often with a husband and wife, one of the debtors may raise a correction to show that the outstanding CCJ is not theirs but their partner's.

CCJs, etc. are information in the public domain. Some county courts can provide the information electronically; in many cases, however, the information will be entered from paper notices and records.

Perhaps the important thing to realize about this category of information—both electoral rolls and CCJs—is that it is all (or almost all) available to the public. The value that the bureaus add is to load it electronically and greatly speed up the process. Thus instead of having to contact several local authorities, which may take several days, if not weeks, one can now access the information in a few seconds.

8.5.2 Previous searches

When a lender makes an inquiry of a credit reference agency, that inquiry is recorded on the consumer's file. (There are special circumstances when it is not recorded, but we do not need to discuss them here.) When another lender makes a subsequent inquiry, a record of any previous searches will be visible. The previous searches carry a date and details of the type of organization that carried it out—bank, insurance company, credit card company, utilities, mail order, etc.

What it does not reveal is the outcome of the inquiry. Therefore, a consumer may have eight searches recorded over a two-week period by a number of companies. We cannot tell which of these are associated with offers from the lenders and in which cases the lender declined the application. For those cases where an offer was made, we cannot tell if the applicant accepted the offer. For example, the consumer could be simply shopping around several lenders for the best deal but will take out only one loan. Or the consumer could be

moving house and the inquiries are to support financing the purchase of furniture, a television, kitchen units, etc. A further possibility is that the consumer is desperately short of money and is applying for a variety of loans or credit cards and intends to take everything that he can get.

Therefore, while the number and pattern of previous searches may be of interest, either in subjective assessment or in a scorecard, it requires careful interpretation. In scorecard development, the characteristics that might be included would be the number of searches in the last 3 months, 6 months, 12 months, and 24 months, as well as the time since the last inquiry. Obviously, with the likely correlation structure, it would be very rare that more than one of these would appear in a final scorecard.

8.5.3 Shared contributed information

Many years ago, both lenders and the bureaus realized that there was value in sharing information on how consumers perform on their accounts. Therefore, at its simplest, several lenders can contribute details of the current performance of their personal loans. If a consumer applies for a personal loan and they currently have one with one of the contributors, an inquiry at the bureau will provide details of whether their existing loan is up-to-date or in arrears and some brief details of the historical payment performance.

This developed in many ways but the fundamental guidelines are encapsulated in a document to which lenders and the bureaus subscribe—the principles of reciprocity. Basically, one gets to see only the same type of information that one contributes. Some lenders contribute data on only some of their products, and they should see details of other lenders' contributions only for the same products. Some lenders do not provide details on all their accounts. Rather, they provide details only of their accounts that have progressed beyond a certain stage in the collections process, usually the issue of a default notice. (Under the Consumer Credit Act of 1974, this is a statutory demand for repayment and precedes legal action.) These lenders—or, more accurately, these lenders when carrying out inquiries relating to this product—will get to see only default information supplied by other lenders. This is the case even if the other lenders also contributed details of all their accounts that are not in serious arrears or in arrears at all.

Although the adult population of the U.K. is only about 40 million, between them the bureaus have approximately 350 million records contributed by lenders. (Note that there will be some duplication and that this number will also include closed or completed accounts for a period after they have ceased to be live.)

When a lender carries out an inquiry, it will not be able to discover with which company the existing facilities and products are. However, it does get to see details of the type of product—revolving credit, mail order, credit card, etc.

The principles of reciprocity not only dictate that you get to see only the same type of information as you contribute. They also dictate the restrictions placed on access to information depending on the purpose to which it will be put. As a rough guideline, the restrictions are least when the information is being used to manage an existing account with an existing customer. Further restrictions are introduced when the data are to be used to target an existing customer for a new product, i.e., one that they do not already hold. Even further restrictions may be placed on the use of shared data for marketing products to noncustomers.

8.5.4 Aggregated information

Through having data from the electoral roll and having the contributed records from many lenders, the bureaus are in an advantageous position to create new measures that might be

of use in credit assessment. With the depth of the information supplied by lenders, together with the electoral roll and the post office records, they do create variables at a postcode level. (Each individual postcode has between 15 and 25 houses allocated to it. In large blocks of flats, there may be one postcode allocated for each block.) Therefore, by aggregating this information, the bureaus are able to create and calculate measures such as

- percentage of houses at the postcode with a CCJ;
- percentage of accounts at the postcode that are up-to-date;
- percentage of accounts at the postcode that are three or more payments in arrears;
- percentage of accounts at the postcode that have been written off in the last 12 months.

Clearly, the individual lender cannot see this information in its entirety as the search is indexed by the address that is entered. However, the bureaus are able to create such measures that, in some scoring developments, prove to be of value. Also, as the ODPR tries to restrict recording and use of data pertaining to the individual, this type of information may be able to restore the lenders' ability to discriminate between good and bad credit risks.

8.5.5 Fraud warnings

The bureaus are able to record and store incidences of fraud against addresses. These could either be first-party fraud, where the perpetrator lives at the address quoted, or impersonation fraud, where the perpetrator claims that they live at the address quoted.

A credit reference inquiry carried out at the address may generate some sort of fraud warning. This does not mean that the application is fraudulent. Most lenders will use this as a trigger to be more careful and perhaps increase the level of checking that is carried out on the details supplied. In fact, many cases that generate a fraud warning are genuine applications; the warning having resulted from an attempted impersonation fraud. It would be counter to common sense, to business sense, and to the guidelines to which the U.K. industry adheres to assume that a fraud warning means the application in fraudulent. While the warning may make the lender more cautious, the lender should decline the case as a fraudulent application only if it has proof that there is some aspect of fraud involved. One cannot decline a case as fraudulent simply because another lender has evidence of an attempted fraud.

8.5.6 Bureau-added value

As can be seen from the above discussion, the credit bureaus store huge amounts of information. Using this, they are able to create and calculate measures. However, they can also develop generic scorecards. Their construction is along the lines of a standard scorecard. There is a huge volume of data and a definition of bad is taken. These scorecards do not relate to a specific lender's experience. Neither do they relate to the lender's market position. They also do not relate to specific products. Further, the bad definition may not be appropriate for a specific situation. However, they can be extremely useful in at least three environments.

The first environment is where the lender is too small to be able to build its own scorecard. To allow it some of the benefits of scoring, the lender calibrates its own experience against the generic scorecard. This will allow it to build some confidence of how the scorecard will operate and also to set a cutoff appropriate to its needs.

The second environment is when there is a new product. In such cases, a generic scorecard may be of use even for larger lenders since there would be insufficient information on which to build a customized scorecard for that product.

The third environment is in general operation. While the lender may have a scorecard that discriminates powerfully between good and bad cases, a generic scorecard allows the lender to do at least two things. The first thing is that the generic scorecard may add to the power of the lender's own scorecard as the generic scorecard will be able to reveal if an applicant is having difficulties with other lenders. The second thing, therefore, is that the generic scorecard can be used to calibrate the quality of applications the lender receives. For example, if the average generic score one month is 5% below the previous month's, the lender has a quasi-independent measure of the credit quality of the recent batch of applications. Because of the different restrictions placed on the use of shared data, each bureau constructs different generic scorecards using different levels of data.

Another example of the development and use of a generic scorecard appears in Leonard (2000). In this example, the credit bureau has built a generic scorecard but has included only cases that are three payments in arrears. Accounts are classified as good if they make payments in the subsequent 90 days. This generic scorecard is intended, therefore, only to discriminate among those accounts already three payments in arrears. It should not be used for accounts that are not in arrears nor for new applications.

In the context of understanding the data that might be available for building a credit scorecard, we have discussed the different types of information. In using credit reference information in the U.K., many other issues need to be borne in mind. However, we do not need to address them here. As we use credit reference information in other countries, other issues will arise.

One issue that we need to discuss is that of matching levels. When carrying out a credit reference inquiry, the lender submits a name and an address. However, in many cases, this does not match exactly with a record in the lender's database. For example, the applicant may have written "S. Jones" on the application form, but the record is for Steven Jones. The address may be written as "The Old Mill, 63 High Street, London," but the record on the database is for 63 High Street, London. Therefore, the bureaus need to have some way to match cases and to return cases that have an exact match, or a very close match, or a possible match. Thus either the lender specifies the level of matching that it requires or the bureaus can return elements of the inquiry with a flag detailing the level of match that was achieved.

Having discussed the information that the bureaus hold and looked at the way they might match details, it should be clear that the credit reference agencies, as they are also called, are just that. Apart from the generic scorecards that they can offer and the way they have aggregated data at postcodes, for all of the other data, they are acting as an agent of the credit industry, accelerating access to available information to support high volumes of applications.

In the U.K., the credit bureaus operate within a legal framework. For example, all consumers have the right, on payment of a small fee, to receive a print of what is held on their files at the credit bureau. The authority to take a credit reference inquiry must be explicitly given by the applicant before the inquiry is made. Also, if the lender wishes to have the authority to take repeated inquiries during the life of the account, this authority must be given in advance, although it can be given once for the life of the account. Repeated inquiries are quite often made in managing credit card limits or in assessing the appropriate action to take when a customer falls into arrears.

The discussion so far has referred to credit reference information on consumers, i.e., on individuals. However, one can also make a credit reference inquiry about a company or business. Once again, the agencies in this market provide accelerated access to public information as well as adding value.

The public information might be details of who the company directors are and with

what other companies they are associated. The financial accounts of the company can also be accessed, as can any court actions, such as CCJs.

The bureaus will also build models to assess the creditworthiness of the business. This is often expressed as a credit limit. In other words, if a business has a credit limit of £10,000, that is the maximum amount that the bureau assesses that creditors should extend to the business. It is not usually possible, however, to find out easily how much credit other creditors have extended to the business toward the credit limit. Bureaus may also assess the strength of the business not only in terms of its financial performance but also in terms of the trends of the relevant industry and economy.

For small businesses, the information available is similar to that for consumers. As we move into larger businesses, the public data become more standardized in their format and in their content. For national and multinational enterprises, the value that a bureau can add diminishes and the bureaus tend toward being simply an accelerated route to public information.

8.6 Determining subpopulations

The next two sections are about deciding which of the variables should be used in the scoring systems and how they should be used. An important use of the variables is to split the population into subpopulations, and different scorecards are built on each subpopulation. There may be policy reasons as well as statistical reasons for doing this. For example, in behavioral scorecards it is usual to build different scorecards for recent customers and for long-existing customers. This is simply because some characteristics, like average balance in the last six months, are available for the latter but not for the former. Another policy might be that the lender wants younger people to be processed differently than older customers. One way to do this is to build separate scorecards for under 25s and over 25s.

The statistical reason for splitting the population into subpopulations is that there are so many interactions between one characteristic and the others that it is sensible to build a separate scorecard for the different attributes of that characteristic. One builds a classification tree using the characteristics that have interactions with lots of other variables. The top splits in such a tree might suggest suitable subpopulations for building separate scorecards. However, it is often found that although there may be strong interactions between one variable and some others, these do not remain as strong when other variables are introduced or removed.

Hence, typically, policy reasons more than statistical ones determine what subpopulations should be formed. As Banasik et al. (1996) showed, segmenting the population does not always guarantee an improved prediction. If the subpopulations are not that different, then the fact that they are built on smaller samples degrades the scorecard more than any advantage the extra flexibility gives. Segmenting the population into two subpopulations depending on whether $X_0 = 0$ or $X_1 = 1$ is equivalent to building a scorecard so that every other variable Y has two variables $Y_0 = (Y|X_0 = 0)$ and $Y_1 = (Y|X_0 = 1)$. On top of this, having two or more scorecards allows more flexibility in the cutoff scores, although one usually tries to ensure that the marginal odds of good to bad is the same at each subpopulation's cutoff score.

8.7 Coarse classifying the characteristics

Once we have divided the population into its subpopulations, one might expect that one could go ahead and apply the methodologies of Chapters 4, 5, and 6 to the variables describing

the characteristics of the applicants. This is not the case. First, one needs to take each characteristic and split the possible answers to it into a relatively small number of classes, i.e., to coarse classify the characteristic. This needs to be done for two different reasons, depending on whether the variable is categorical (has a discrete set of answers) or continuous (has an infinite set of possible answers). For categorical characteristics, the reason is that there may be too many different answers or attributes, and so there may not be enough of the sample with a particular answer to make the analysis robust. For continuous characteristics, the reason is that credit scoring seeks to predict risk rather than to explain it, and so one would prefer to end up with a system in which the risk is nonlinear in the continuous variable if that is a better prediction. The following examples show these two effects.

Example 8.1 (residential status). Suppose that the distribution of answers among a sample of 10,000 to the question "What is your residential status?" was as given in Table 8.2. The overall good:bad ratio in the population is 9:1, but the ratios in the attributes range from 20:1 to 1:1. The question is whether we can allow all six types of answer (attribute) to remain. Only 20 of the populations gave "No answer"; 140 were "Other," while "Rent furnished" had only 5% of the population in that category. There are no hard and fast rules, but it would seem that these categories have too few answers for there to be much confidence that their good:bad ratios will be reproduced in the whole population. We could put all three in one category—"Other answers," with 450 goods and 200 bads—since they are the three worst good:bad ratios. There is an argument though for putting rent furnished with rent unfurnished since again their good:bad odds are not too far apart and they are both renting categories. Similarly, if we were to decide on only three categories, should it be (owner/renter) owners (6000 good, 300 bad), renters (1950 goods, 540 bads), and others (1050 good, 160 bad) or should it be (owner/parent) consisting of owners (6000 good, 300 bad), with parents (950 good, 100 bad), and others (2050 good, 600 bad).

Table 8.2. *Table of numbers in the different groups.*

Attribute	Owner	Rent unfurnished	Rent furnished	With parents	Other	No answer
Goods	6000	1600	350	950	90	10
Bads	300	400	140	100	50	10
Good:bad odds	20:1	4:1	2.5:1	9.5:1	1.8:1	1:1

Although we said that the choice of combination is as much art as it is science, one can use some statistics as guidance. Three statistics are commonly used to describe how good the characteristic with a particular coarse classification is at differentiating goods from bads. The most common is the χ^2-statistic.

8.7.1 χ^2-statistic

Let g_i and b_i be the numbers of goods and bads with attribute i, and let g and b be the total number of goods and bads. Let

$$\hat{g}_i = \frac{(g_i + b_i)g}{g + b} \quad \text{and} \quad \hat{b}_i = \frac{(g_i + b_i)b}{g + b}$$

be the expected number of goods and bads with attribute i if the ratio for this attribute is the same as for the whole population. Then

$$s^2 = \sum_i \left(\frac{(g_i - \hat{g}_i)^2}{\hat{g}_i} + \frac{(b_i - \hat{b}_i)^2}{\hat{b}_i} \right) \tag{8.1}$$

is the χ^2-statistic. Formally, this measures how likely it is that there is no difference in the good:bad ratio in the different classes, and one can compare it with the χ^2-statistics with $k-1$ degrees of freedom, where k is the number of classes of the characteristic. However, one can use it as a measure of how different the odds are in the different classes with a higher value, reflecting greater differences in the odds. Thus in the three class cases above, we get

$$\begin{aligned}
\text{(owner/renter):}\quad & \hat{g}_{\text{owner}} = 5670, \quad \hat{g}_{\text{renter}} = 2241, \quad \hat{g}_{\text{others}} = 1089, \quad \text{so} \\
\chi^2 &= \frac{(6000-5670)^2}{5670} + \frac{(300-630)^2}{630} + \frac{(1950-2241)^2}{2241} \\
&\quad + \frac{(540-249)^2}{249} + \frac{(1050-1089)^2}{1089} + \frac{(160-121)^2}{121} \\
&= 583.9, \qquad (8.2) \\
\text{(owner/parent):}\quad & \hat{g}_{\text{owner}} = 5670, \quad \hat{g}_{\text{parent}} = 945, \quad \hat{g}_{\text{others}} = 2385, \quad \text{so} \\
\chi^2 &= \frac{(6000-5670)^2}{5670} + \frac{(300-630)^2}{630} + \frac{(950-945)^2}{945} \\
&\quad + \frac{(100-105)^2}{105} + \frac{(2050-2385)^2}{2385} + \frac{(600-265)^2}{265} \\
&= 662.9. \qquad (8.3)
\end{aligned}$$

One would say the owner/parent split with its larger χ^2 value is the better split.

8.7.2 Information statistic

The information statistic is related to measures of entropy that appear in information theory and is defined by

$$F = \sum_i \left(\frac{g_i}{g} - \frac{b_i}{b} \right) \log \left[\frac{g_i b}{b_i g} \right]. \tag{8.4}$$

In statistics, this is sometimes called the divergence or information value. It seeks to identify how different $p(x|G)$ and $p(x|B)$ are (which translate into $\frac{g_i}{g}$ and $\frac{b_i}{b}$) when x takes attribute value i. The motivation of the information statistic being of the form

$$\sum_i \left(\frac{g_i}{g} \right) \log \left[\frac{g_i}{g} \right]$$

when there are g observations and g_i of them were of type i is as follows. The number of ways this distribution occurs is $N_g = \frac{g!}{g_1! g_2! \ldots g_p!}$ if there are p types in total. Information is taken as the log of the number of different ways one can get the message that was seen, so $I_g = \log N_g = \log g! - \sum_i \log(g_i!) \approx g \log(g) - \sum_i g_i \log(g_i)$. The average information is $\frac{I_g}{g} \approx -\sum_i (\frac{g_i}{g})(\log(g_i) - \log(g)) = -\sum_i (\frac{g_i}{g}) \log(\frac{g_i}{g})$. This information statistic can be thought of as the difference between the information in the goods and the information in the bads, i.e., $-\sum_i (\frac{g_i}{g} - \frac{b_i}{b})(\log(\frac{g_i}{g}) - \log(\frac{b_i}{b}))$.

In Example 8.1, we get

$$\text{(owner/renter):} \quad \frac{g_{\text{owner}}}{g} = 0.667, \quad \frac{b_{\text{owner}}}{b} = 0.3, \quad \frac{g_{\text{renter}}}{g} = 0.217, \quad \frac{b_{\text{renter}}}{b} = 0.54,$$
$$\frac{g_{\text{other}}}{g} = 0.117, \quad \text{and} \quad \frac{b_{\text{owner}}}{b} = 0.16,$$
$$F = (0.667 - 0.3)\log\left(\frac{0.667}{0.3}\right) + (0.217 - 0.54)\log\left(\frac{0.217}{0.54}\right)$$
$$+ (0.117 - 0.16)\log\left(\frac{0.117}{0.16}\right)$$
$$= 0.6017, \tag{8.5}$$

$$\text{(owner/parent):} \quad \frac{g_{\text{owner}}}{g} = 0.667, \quad \frac{b_{\text{owner}}}{b} = 0.3, \quad \frac{g_{\text{parent}}}{g} = 0.106, \quad \frac{b_{\text{parent}}}{b} = 0.1,$$
$$\frac{g_{\text{other}}}{g} = 0.228, \quad \text{and} \quad \frac{b_{\text{owner}}}{b} = 0.6,$$
$$F = (0.667 - 0.3)\log\left(\frac{0.667}{0.3}\right) + (0.106 - 0.1)\log\left(\frac{0.106}{0.1}\right)$$
$$+ (0.228 - 0.6)\log\left(\frac{0.228}{0.6}\right)$$
$$= 0.6536. \tag{8.6}$$

Large values of F arise from large differences between $p(x|G)$ and $p(x|B)$ and so correspond to characteristics that are more useful in differentiating goods from bads. Thus in this case, we would take the (owner/parent) split again.

8.7.3 Somer's D concordance statistic

The Somer's D concordance statistic test assumes that the classes of the characteristic already have an ordering from low, which has the lowest good rate, to high, which has the highest good rate. The concordance statistic describes the chance that if one picks a good at random from the goods and a bad at random from the bads, the bad's attribute, x_B, will be in a lower class than the good's attribute, x_G. The higher this probability, the better the ordering of the characteristic's classes reflects the good-bad split in the population. The precise definition of D is the expected payoff of a variable, which is 1 if the ordering puts the bad below the good, -1 if it puts the good below the bad, and 0 if they are the same:

$$D = 1.P\{x_B < x_G\} - 1.P\{x_B > x_G\} + 0.P\{x_B = x_G\}$$
$$= \sum_i \frac{\left(\sum_{j<i} b_j\right) g_i - \left(\sum_{j<i} g_j\right) b_i}{bg}. \tag{8.7}$$

For the three-class example of Example 8.1, we get the following calculations: For (owner/renter), the ordering is that renters have the lowest good rate, then others, then owners, so

$$D = \left(\frac{540}{1000}\right)\cdot\left(\frac{1050}{9000}\right) + \left(\frac{700}{1000}\right)\cdot\left(\frac{6000}{9000}\right) - \left(\frac{1950}{9000}\right)\cdot\left(\frac{160}{1000}\right) - \left(\frac{3000}{9000}\right)\cdot\left(\frac{300}{1000}\right)$$
$$= 0.395. \tag{8.8}$$

For the (owner/parent) split, the lowest good rate is others, then parent, and then owners. Hence

$$D = \left(\frac{600}{1000}\right)\cdot\left(\frac{950}{9000}\right)+\left(\frac{700}{1000}\right)\cdot\left(\frac{6000}{9000}\right)-\left(\frac{2050}{9000}\right)\cdot\left(\frac{100}{1000}\right)-\left(\frac{3000}{9000}\right)\cdot\left(\frac{300}{1000}\right)$$
$$= 0.4072. \tag{8.9}$$

Since (owner/parent) has the higher value, this test suggests that it is a more definitive split.

Thus far, we have dealt with categorical variables, but we also need to coarse classify continuous variables. The first question is why? In normal regressions, if one has a continuous variable, one usually leaves it as such and calculates only the coefficient that it needs to be weighted by. That is because one is trying to explain some connection. If instead one is trying to predict risk, then by leaving the variable continuous, one guarantees the risk will be monotone in that variable. For example, Figure 8.1 shows the good rate at each age group from a sample of credit card holders. What is noticeable is that it is not monotone—the good rate goes up, then down, and then up again as age progresses. One can find explanations for this in terms of more responsibility and more outgoings in the 30–40 age group or the loss of a second income during the childbearing years, but the point is that the best regression line (shown in Figure 8.1) does not reflect this. It is therefore better to split age into a number of distinct attributes, 18–21, 21–28, 29–36, 37–59, 60+, and as Figure 8.2 shows, these can have scores that reflect the nonlinearity with age.

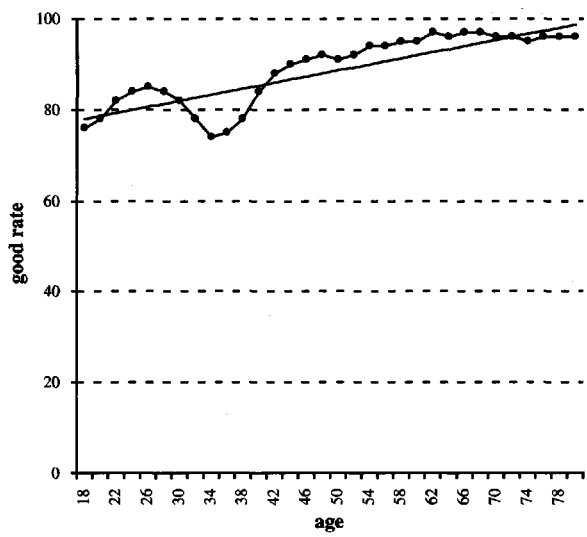


Figure 8.1. *Good rate as a function of age.*

So how should one split such continuous variables? The most obvious way is to split the characteristics first into attributes corresponding to percentiles, i.e., into 10 classes, the first having the youngest 10%, the second class the second youngest 10%, etc. There is no need to fix on 10. Some analysts use 20 groups, each of 5% of the population, or eight groups of 12.5% of the population, or even 100 groups of 1% each. Once this is done, the question is whether one should group nearest neighbors together. This depends on how close their good rates are, and one can use the statistics outlined above to help decide. Let us take time at present address as an example.

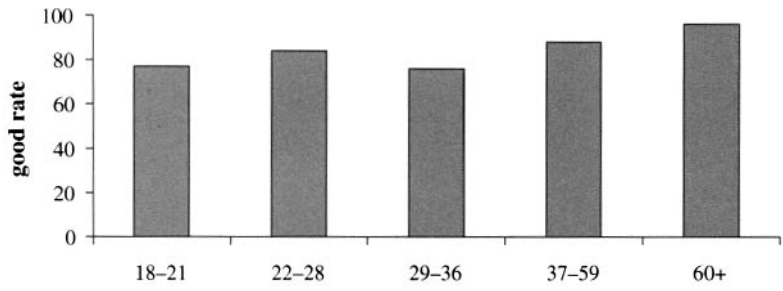


Figure 8.2. *Coarse classifying of age.*

Example 8.2 (time at present address). The data (from a hypothetical example) is given in Table 8.3, where TPA is time at present address. Choosing splits by inspection, first one might decide that the five groups of more than four years have such high good rates that they can be put together. The nonlinearities at the short times at address mean that one might keep the other five groups separate, but suppose that we want to put together at least one more pair. Some might argue for putting the under-6 months and the 6–12 months together in a <12 months group (option 1), while others might want to put 18–30 months and 30–48 months together since their ratios are close (option 2). One would not put the 12–18 months and the 30–48 months groups together because although their ratios are very similar, they are not adjacent time periods. To aid this choice between the two options, one could calculate the statistics as before.

Table 8.3. *Data for Example* 8.2.

TPA	<6 months	6–12 months	12–18 months	18–30 months	30–48 months	4–5 years	6–7 years	8–11 years	12–15 years	16+ years
Goods	800	780	840	880	860	920	970	980	980	990
Bads	200	220	160	120	140	80	30	20	20	10
Total	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
Ratio	4:1	3.5:1	5.2:1	7.3:1	6.1:1	11.5:1	31:1	49:1	49:1	99:1

Take option 1 with groups of <12 months (1580 goods, 420 bads), 12–18 months (840 goods, 160 bads), 18–30 months (880 goods, 120 bads), 30–48 months (860 goods, 140 bads), and over four years (4840 goods, 160 bads). Let option 2 be <6 months (800 goods, 200 bads), 6–12 months (780 goods, 220 bads), 12–18 months (840 goods, 160 bads), 18–48 months (1740 goods, 260 bads), and over four years (4840 goods, 160 bads). The calculations are as follows.

χ^2 statistic:

$$\begin{aligned}\text{Option 1:}\quad \chi^2 &= \frac{(1580-1800)^2}{1800} + \frac{(840-900)^2}{900} + \frac{(880-900)^2}{900} + \frac{(860-900)^2}{900} \\ &\quad + \frac{(4840-4500)^2}{4500} + \frac{(420-200)^2}{200} + \frac{(160-100)^2}{100} \\ &\quad + \frac{(120-100)^2}{100} + \frac{(140-100)^2}{100} + \frac{(160-500)^2}{500} \\ &= 588. \end{aligned} \tag{8.10}$$

$$\text{Option 2:}\quad \chi^2 = \frac{(800-900)^2}{900} + \frac{(780-900)^2}{900} + \frac{(840-900)^2}{900} + \frac{(1740-800)^2}{800} + \frac{(4840-4500)^2}{4500} + \frac{(200-100)^2}{100} + \frac{(220-100)^2}{100} + \frac{(160-100)^2}{100} + \frac{(260-200)^2}{200} + \frac{(160-500)^2}{500} = 588. \tag{8.11}$$

So options 1 and 2 are considered equally good under this test.

Information statistic F:

$$\text{Option 1:}\quad F = \left(\frac{1580}{9000} - \frac{420}{1000}\right)\log\left(\frac{1580\cdot 1000}{9000\cdot 420}\right) + \left(\frac{840}{9000} - \frac{160}{1000}\right)\log\left(\frac{840\cdot 1000}{9000\cdot 160}\right) + \left(\frac{880}{9000} - \frac{120}{1000}\right)\log\left(\frac{880\cdot 1000}{9000\cdot 120}\right) + \left(\frac{860}{9000} - \frac{140}{1000}\right)\log\left(\frac{860\cdot 1000}{9000\cdot 140}\right) + \left(\frac{4840}{9000} - \frac{160}{1000}\right)\log\left(\frac{4840\cdot 1000}{9000\cdot 160}\right) = 0.7287. \tag{8.12}$$

$$\text{Option 2:}\quad F = \left(\frac{800}{9000} - \frac{200}{1000}\right)\log\left(\frac{800\cdot 1000}{9000\cdot 200}\right) + \left(\frac{780}{9000} - \frac{220}{1000}\right)\log\left(\frac{780\cdot 1000}{9000\cdot 220}\right) + \left(\frac{840}{9000} - \frac{160}{1000}\right)\log\left(\frac{840\cdot 1000}{9000\cdot 160}\right) + \left(\frac{1740}{9000} - \frac{260}{1000}\right)\log\left(\frac{1740\cdot 1000}{9000\cdot 260}\right) + \left(\frac{4840}{9000} - \frac{160}{1000}\right)\log\left(\frac{4840\cdot 1000}{9000\cdot 160}\right) = 0.7280. \tag{8.13}$$

Thus in this case, the suggestion is to take the larger value—namely, option 1—but it is very close.

***D* concordance statistic:**

$$\text{Option 1:}\quad D = \left(\frac{420}{1000}\right)\left(\frac{840}{9000}\right) + \left(\frac{580}{1000}\right)\left(\frac{860}{9000}\right) + \left(\frac{720}{1000}\right)\left(\frac{880}{9000}\right) + \left(\frac{840}{1000}\right)\left(\frac{4840}{9000}\right) - \left(\frac{1580}{9000}\right)\left(\frac{160}{1000}\right) - \left(\frac{2420}{9000}\right)\left(\frac{140}{1000}\right) - \left(\frac{3280}{9000}\right)\left(\frac{120}{1000}\right) - \left(\frac{4160}{9000}\right)\left(\frac{160}{1000}\right) = 0.433. \tag{8.14}$$

$$\text{Option 2:}\quad D = \left(\frac{220}{1000}\right)\left(\frac{800}{9000}\right) + \left(\frac{420}{1000}\right)\left(\frac{840}{9000}\right) + \left(\frac{580}{1000}\right)\left(\frac{1740}{9000}\right) + \left(\frac{840}{1000}\right)\left(\frac{4840}{9000}\right) - \left(\frac{780}{9000}\right)\left(\frac{200}{1000}\right) - \left(\frac{1580}{9000}\right)\left(\frac{160}{1000}\right) - \left(\frac{2420}{9000}\right)\left(\frac{260}{1000}\right) - \left(\frac{4160}{9000}\right)\left(\frac{160}{1000}\right) = 0.433. \tag{8.15}$$

Thus both options have the same value under this criterion.

Note that different options are chosen by the different statistics.

Although we described that one of the reasons for coarse classifying continuous variables was to model the situation when the good rate is nonmonotonic, there may be cases in which the lenders want the good rate in a particular characteristic to be monotone if not necessarily linear. This may be because the lenders have prior beliefs about the effect of income, or time at bank on the good rate, or because they want to bias the scorecards to one or other end of the characteristic—for example, toward younger or older customers. In that case, we use the following coarse classification approach.

8.7.4 Maximum likelihood monotone coarse classifier

Assume that the bad rate is going down as the characteristic value increases. (If not, apply the following procedure starting from the other extreme of the characteristic.) Start at the lowest characteristic value and keep adding values until the cumulative bad rate hits its maximum. This is the first coarse classification split point. Start calculating the cumulative bad rate from this point until it again hits a maximum. This is the second split point. Repeat the process until all the split points are obtained.

It is an interesting exercise to show that, in fact, this algorithm gives the maximum-likelihood estimate of the fit conditional on the bad rates decreasing from class to class. That is a subtle way of saying that we will not prove it in this text but leave it for the inquiring reader to prove. We apply this procedure to the following two examples.

Example 8.2 (revisited). Using the data of Example 8.2, the calculations in Table 8.4 find the first four split points. Thus the classes are <12 months, 12–18 months, 18–48 months, and 4–5 years; doing the rest of the calculations shows that all the remaining classes remain separate, except that 8–11 years and 12–15 years clearly can be put together.

Table 8.4. *Calculations for Example* 8.2 *(revisited).*

TPA	<6 months	6–12 months	12–18 months	18–30 months	30–48 months	4–5 years	6–7 years	8–11 years	12–15 years	16+ years
Goods	800	780	840	880	860	920	970	980	980	990
Bads	200	220	160	120	140	80	30	20	20	10
Cum. bad rate	0.2	0.21*	0.193	0.175						
Cum. bad rate			0.16*	0.14	0.13	0.125				
Cum. bad rate				0.12	0.13*	0.113				
Cum. bad rate						0.0*	0.055	etc.		

The procedure can also be applied at an individual data level as follows.

Example 8.3. In a sample of consumers, the consumers have the values of a particular continuous characteristic, as shown in Table 8.5, and their good bad status is denoted by G or B. We apply the procedure described at the start of the section, data point by data point. Thus the classes are <12 (bad rate 0.71), 12–17 (bad rate 0.67), 18–25 (bad rate 0.6), 26–36 (bad rate 0.57), and 36+ (bad rate 0).

Coarse classifying characteristics is essentially the same thing as splitting the population into subclasses, which is the essence of the classification tree approach to classification. Therefore, one could also use the Kolmogorov–Smirnov statistics or indices like the Gini

Table 8.5. *Calculations for Example* 8.3.

Character value	1	3	6	7	9	10	12	14	16	17	18	20
Good-bad	G	B	B	G	B	B	B	G	B	B	G	B
Cumulative bad	0	.5	.67	.5	.6	.67	.71*	.62	.67	.7	.63	.67
								0	.5	.67*	.5	.6
											0	.5

Character value	21	24	25	27	29	30	31	32	34	36	37	38
Good-bad	G	B	B	G	B	G	B	G	B	B	G	G
Cumulative bad												
	.5	.55	.62	.55	.6	.54						
	.33	.5	.60*	.5	.57	.5						
				0	.5	.33	.5	.4	.5	.57*	.5	.44

index, which were used in section 4.7 to decide on the best splits in the trees, as a way to decide how to coarse classify a variable. The difference between coarse classifying and tree splitting is one of timing rather than approach. If one takes a variable and splits it into classes expecting thereafter to use some other scoring approach like regression or linear programming, then it is coarse classifying. If one splits the variable as part of the classifying process, then it is considered to be the splitting part of the tree-building process.

8.8 Choosing characteristics

When the variables have been coarse classified, one can end up with a large number of attributes. For application scorecards with, say, 30 to 40 variables, initially one could have more than 200 attributes, while for behavioral scorecards, which sometimes have hundreds of initial characteristics, one could be dealing with more than 1000 different attributes. If one were to take each attribute as a binary variable, this would be far too many variables for a logistic regression or a classification tree to cope with sensibly. Moreover, if one wants to end up with a scorecard that is understandable and hence accepted by managers, it should not have much more than 20 characteristics in it. So how should these characteristics be chosen?

One can rule out some variables because they are poor predictors, too variable over time, or too dependent on other variables. The first can be checked using the statistics used in the last section to decide on the coarse classification. If one calculates the χ^2-statistic, the information statistic F, or the concordance statistic D for each characteristic, these statistics give crude orderings of the importance of the variables as predictors.

Consider the following example.

Example 8.4. Consider the characteristics X_1, X_2, and X_3, all of which are binary variables. Table 8.6 gives the data and the values of the various statistics. This suggests that one would choose X_1 rather than X_2 or X_3. It is preferred to X_2 because it has more discrimination and to X_3 because there is a more even split between the numbers in the different characteristics. However, χ^2 does not show this, while F and D do not pick up the difference between X_2 and X_3 in terms of the size of the populations splits.

Blackwell (1993) pointed out the limitations in these characteristics in that the χ^2 test is looking at the statistical significant difference from all attributes having the same bad rate, which is not the same as discriminatory power. The information statistic F, on the other hand, does measure discriminatory power well, but it is very insensitive to sample size.

Table 8.6. *Data and calculations of Example* 8.4.

	X_1		X_2		X_3	
	Good	**Bad**	**Good**	**Bad**	**Good**	**Bad**
$X_i = 1$	2000	3000	4000	3000	5800	3600
$X_i = 0$	4000	1000	2000	1000	200	400
χ^2	1.611		5.480		245.9*	
F	0.746*		0.0337		0.078	
D	0.416*		0.083		0.067	

Blackwell suggests an efficiency measure $H(p)$, where p is the decision rule that accepts those whose probability of being good exceeds p. $H(p)$ takes into account $\frac{D}{L}$, the relative loss of accepting a bad to rejecting a good. He defines $g(p)$ to be the number of goods in attributes of the characteristic where the good rate is larger than p and defines $b(p)$ to be the number of bads in attributes of the characteristic where the good rate is below p. Let g and b be the total numbers of good and bad in the population; then

$$H(p) = \begin{cases} \dfrac{Lg(p) + Db(p) - Lg}{Db} & \text{if } \dfrac{g}{g+b} > p, \\ \dfrac{Lg(p) + Db(p) - Db}{Lg} & \text{if } \dfrac{g}{g+b} \leq p. \end{cases} \tag{8.16}$$

If a characteristic can be split into attributes that are either all good or all bad, $g(p) = g$ and $b(p) = b$, so $H(p) = 1$, while if all the attributes have the same proportion of goods, then if $\frac{g}{g+b} > p$, $g(p) = g$ and $b(p) = 0$; so $H(p) = 0$, with a similar result if $\frac{g}{g+b} \leq p$. If one wanted to use this to rank characteristics, one would have to decide on a specific value of p or else integrate $H(p)$ over all p, i.e., $\int H(p)dp$. However, it is more usual to use χ^2, F, and D to measure the discrimination of the characteristics.

One can check for the robustness of a characteristic over time by splitting the sample into subsamples depending on the date the account was opened and checking whether the distribution of the characteristic was consistent over these time-dependent cohorts of accounts. Dependence can be checked using the correlation between variables, but common sense also plays its part. If, for example, there are several income variables in an application score data set or several variants of the average balance variables in a behavioral score data set, it is likely that the scorecard will want to have only one of the group in it. One way to choose is to run a regression of the good-bad variable on this group of variables and chose which one first enters the regression under forward selection or which one is the last one left under backward selection.

If one is using dummy variables to define the characteristics after coarse classification, then this stepwise linear-regression approach is also a good way to cut down the number of variables at that stage. The advantage of using linear regression over the other techniques, including logistic regression, at this choice stage is that it is much faster to run and there could be lots of variables still around. It also produces more poorly aligned scorecards, and this makes it more obvious which attributes are unsatisfactory predictors.

An alternative approach that can be used after performing the coarse classifying of the previous section does not increase the number of variables from the original number of characteristics. In this approach, we replace the value of attribute i of a characteristic by the good:bad odds or the weight of evidence for that attribute. This has the advantage of not increasing the number of variables at all and so is particularly useful for the regression

approaches. It ensures that, provided the coefficient of the modified variable is positive, the scores for the attributes will reflect the ranking of their good:bad odds. The problem with it is that there is a bias because these modified independent variables are now not independent of the dependent good:bad variable in the regression. This could be overcome by calculating the good:bad odds or the weight of evidence on part of the sample and calculating the regression coefficients on the remainder. However, this is rarely done in practice. It is also the case that since setting the weight of evidence attribute score is a univariate exercise, the results will be poor if there are still strong correlations between the variables. It certainly is a way to cut down on the number of variables produced, but even in this case one would still use stepwise, forward, or backward regression to remove some of the variables.

8.9 Reject inference

One of the major problems in developing scoring models in credit granting is that only the accounts that were approved in the past have performance data, which allows them to be classified as good or bad. For the customers that were declined in the past, one only has their characteristic values but not their good-bad status. If these customers are ignored and dropped out of the sample, then it no longer reflects the through-the-door population that one is after. This causes a bias (the "reject bias") in any classification procedure built on this sample. A number of techniques have been proposed to overcome this bias, and they come under the name of reject inference. How can one use the partial information available on the rejects to improve the scoring system? Since this has been an area which scorecard builders have sought commercial advantage, it is not surprising that it is an area of some controversy.

The cleanest way to deal with reject bias is to construct a sample in which no one was rejected. Retailers and mail-order firms traditionally do this. They take everyone who applies in a certain period with the intention of using that sample for building the next generation of scorecards. The culture in financial organizations does not accept this solution. "There are bads out there and we just cannot take them" is the argument—although, of course, the greater loss involved in someone going bad on a personal loan, compared with not paying for the two books ordered from a book mail-order club, may have something to do with it. However, this approach has considerable merit and can be modified to address the concerns that it will cost too much in losses. Traditionally banks take probability of default as their criterion because they assume that the losses from defaulters do not have huge variations. In that case, one can diminish these losses by not taking everyone but taking a proportion $p(x)$ of those whose probability of defaulting is x. One lets this proportion vary with x; it is very small when x is almost 1, and it tends to 1 when x is small. By allowing the possibility of picking everyone, reweighting would allow one to reconstruct a sample with no rejection without having to incur the full losses such a sample brings. Although this approach is not finding much support among credit granters, some are trying to get information on reject performance from other credit granters who did give these consumers credit.

If one has rejects in the sample on which the system is to be built, then there are five ways that have been suggested for dealing with rejects: define as bads, extrapolation, the industry norm—augmentation, mixed populations, and the three-way group approach.

8.9.1 Define as bad

The crudest approach is to assign bad status to all the rejects on the grounds that there must have been bad information about them for them to have been rejected previously. The

scoring system is then built using this full classification. The problems with this approach are obvious. It reinforces the prejudices of bad decisions of the past. Once some group of potential customers has been given a bad classification, no matter how erroneously, they will never again get the opportunity to disprove this assumption. It is a wrong approach on statistical grounds and wrong on ethical grounds.

8.9.2 Extrapolation

Hand and Henley (1993), in their careful analysis of reject inference, pointed out that two different situations can occur depending on the relationship between the characteristics X_{old} of the system, which was used for the accept-reject decision, and X_{new}, the characteristics available to build the new scorecard. If X_{old} is a subset of X_{new}, i.e., the new characteristics include all those that were used in the original classification, then for some combination of attributes (those where X_{old} rejected the applicant), we will know nothing about the good-bad split because they were all rejected. For the other combinations, where X_{old} accepted the applicants, we should have information about the proportion of goods to bads but only among the accepted ones. However, X_{old} will have accepted all the applicants with these combination of characteristics. Then one has to extrapolate, i.e., fit a model for all the probability of being good for the attribute combinations that were accepted and extend this model to the combinations that were previously rejected. As was pointed out forcibly by Hand and Henley (1993), this method works far better on methods that estimate $q(G|x)$ directly, like logistic regression, rather than ones that estimate $p(x|G)$ and $p(x|B)$, like linear regression–discriminant analysis. This is because when estimating $p(x|G)$, the sampling fraction is varying with x and so will lead to biases when one tries to estimate the parameters of $p(x|G)$ (e.g., the mean and variance of the normal distribution in discriminant analysis). However, for $q(G|x)$, the fraction of the underlying population with that value of x which is sampled is either 0 or 1, so there is no bias in the model's parameter estimation. What happens is that the model gives a probability of being good to each of the population that was rejected and the scoring systems are then built on the whole population with the rejects having this value. This would not work for methods like nearest neighbors, but it might work for logistic regression if one can believe the form of the model.

8.9.3 Augmentation

If X_{old} is not a subset of X_{new}, so there were unknown variables or reasons for the original rejection decisions, then the situation is even more complicated. The usual approach is the augmentation method, outlined by Hsia (1978). First, one builds a good-bad model using only the accepted population to estimate $p(G|x, A)$, the probability of being good if accepted and with characteristic values x. One then builds an accept-reject model using similar techniques to obtain $p(A|x) = p(A|s(x)) = p(A|s)$, where s is the accept-reject score. The original approach of Hsia (1978) then makes the assumption that $p(G|s, R) = p(G|s, A)$—the probability of being good is the same among the accepteds and the rejecteds at the same accept-reject score, where

$$p(G|s, A) = \sum_{x;s(x)=0} p(G|x, A)p(x|s(x) = s). \tag{8.17}$$

This is like reweighting the distribution of the sample populations so that the percentage with a score s moves from $p(A, s)$ to $p(s)$. A new good-bad scorecard is now built on

the full sample including these rejects. The rejects with accept-reject score s are given the probability $p(G|s, A)$ of being good.

Other methodologies for assessing the $p(G|s, R)$ have also been suggested. One approach assumes that $p(G|s, R) \leq p(G|s, A)$ and chooses this probability subjectively. The discount could depend on the type of account, whether there have been CCJs against the person, and when the account was opened. Others have assumed that $p(G|s, R) = kp(G|s, A)$, where k might be obtained by bootstrapping using a subset of the variables to build a good-bad score, which is then used to obtain the accept-reject decision. However, all these variants of augmentation have some strong assumptions in them concerning the form of the distributions or the relationship between $p(G|s, R)$ and $p(G|s, A)$. In practical applications, these assumptions are not validated and are rarely true.

8.9.4 Mixture of distributions

If one is making assumptions, an alternative approach is to say that the population is a mixture of two distributions—one for the goods and one for the bads—and that the form of these distributions is known. For example, if $p(\mathbf{x})$ is the proportion of applicants with characteristic $\mathbf{x}$, one says that

$$p(\mathbf{x}) = p(\mathbf{x}|G)p_G + p(\mathbf{x}|B)p_B, \tag{8.18}$$

and one can estimate the parameters of $p(\mathbf{x}|G)$ and $p(\mathbf{x}|B)$ using the accepts and, by using the EM algorithm, even the rejects. A usual assumption is that $p(\mathbf{x}|G)$ and $p(\mathbf{x}|B)$ are multivariate normal despite the fact that so many characteristics are categorical or binary. A halfway approach between this approach and augmentation is to assume that the good-bad scores and the accept-reject score for the goods are a bivariate normal distribution with a similar assumption for the bads. In this approach, one initially estimates the parameters of these distributions from the accepts and uses these parameters to estimate the probability of each reject being a good. Using these new estimates for the probability of a reject being a good, reestimate the parameters of the two distributions and iterate the process until there is convergence.

8.9.5 Three-group approach

A final approach is to classify the sample into three groups: goods, bads, and rejects. This was proposed by Reichert, Cho, and Wagner (1983), but the problem is that we want to use our classification system to split future applicants into only two classes—the goods, whom we will accept, and the bads, whom we will reject. What one does with those who are classified as rejects is not clear. If one rejects them, this approach reduces to classifying all rejects as bad. Its only saving grace is that in classical linear discriminant analysis, when one classifies into three groups, it is assumed that all three groups have a common covariance matrix. Hence this is a way to use the information in the rejects to improve the estimation of the covariance matrix.

To sum up, it does seem that reject inference is valid only if one can confidently make some assumptions about the accepted and rejected populations. It may work in practice because these assumptions may sometimes be reasonable or at least moving in the right direction. For example, it must be sensible to assume that $p(G|s, R) < p(G|s, A)$ even if one cannot correctly estimate the drop in probability.

In all the above approaches, we have concentrated on the rejects—the customers to whom the lender decided not to give credit—but exactly the same problem (and hence the same procedures) would need to be applied to deal with the bias of these who withdraw after being made the offer of a loan. Withdrawal might mean exactly that in the case of personal loans, or it might mean never using the loan facility in the case of credit cards.

8.10 Overrides and their effect in the scorecards

Overrides are when the lender decides to take an action contrary to the recommendation of the scoring system. High-side overrides (HSOs) are when the applicant is not given a loan although the score is above the cutoff; low-side overrides (LSOs) are when the applicant is accepted for a loan although the score is below the cutoff. Such overrides can be because the lender has more information than that in the scorecard or because of company policy, or it may be a subjective decision by the credit manager or underwriter.

Informational overrides are usually rare, but it could happen that an accompanying letter or knowledge by the branch suggests one of the characteristics in the scorecard does not tell the whole story. A raise in salary has been authorized but not paid, for example. The facility should be available for the decision to be reversed in such a case.

Policy overrides are when the lender has decided that certain types of applicant will be treated in certain ways irrespective of their other characteristics. For example, it may be decided that students are good long-term prospects and so should be given an overdraft no matter what. It may also be politic for all employees of the company to be accepted for the credit facility. An apocryphal story says that when a major international retailer first introduced store cards, the scorecard rejected the applications from the wives of its directors. An override was immediately instituted.

Many LSOs are justified because the possible loss of lucrative business if the applicant is upset by being rejected. The same argument is used when third parties (mortgage brokers, for example) demand that a loan for one of their clients be agreed. It is important to have the information to make a sensible decision in such cases. How much business and, more important, how much profit could be lost, and how likely is it that the loan will default? One needs to weigh these factors against one another. This means that it should be standard policy for the lender to keep track of the subsequent performance of overrides so that these estimates of losses can be calculated. Sangha (1999) looked in some detail at what information needs to be collected and the pitfalls, including possible noncompliance with the fair-lending laws, that need to be watched. One obvious point is to ensure that if a policy override is known, then those affected by the override should be removed from the sample on which the scoring system is developed.

When underwriters or credit evaluators are running the scoring system, subjective overrides are common but unjustifiable. Either the underwriter or the scoring system is wrong. If it is the underwriter, then he should not have interfered; if it is the scoring system, then the underwriter obviously has generic information that can improve the scorecard and so the scorecard should be redeveloped to include this. The way to find out is again to keep a record of how such overrides perform and get whoever is wrong to learn from this experience. The problem usually is human nature. An underwriter or credit evaluator wants to be seen as useful to protect his position, and accepting all the decisions of an automatic system does not do this. The trick is to redefine the evaluator's role into one of verifying and validating the system, i.e., checking that it is doing what it is supposed to be doing and that the population being assessed is still sufficiently close to the population the system was built on for the decisions still to be sensible.

8.11 Setting the cutoff

When introducing a new scorecard, there are several ways to select a cutoff.

A simplistic approach, especially for the first period following live implementation, is to adopt a cutoff that produces the same acceptance rate as the existing scorecard (on the current quality of applicants). This can be adopted until we have more confidence in the scorecard and its programming and in the other implementation issues that may arise. Even if the objective of the new scorecard is to increase acceptance and maintain the bad rate, it might still be worthwhile to retain the current acceptance rate for a few weeks at least. This can instill confidence in the programming and implementation, and it allows one to flush out any anomalies without many things being changed. Anomalies that might occur would be in the swap sets—those cases that previously were declined and are now approved, and vice versa. For example, we may have had an old scorecard where the acceptance rates were 74% of existing branch customers and 53% for noncustomers. With the new scorecard, these percentages may be 66% and 58%. If we have tried to have the same acceptance rate, then the cause of this shift is either a change in the relative quality of applicants from these two sources or the change in scorecard. If at the same time we have tried to increase acceptance rates and also reduce our expected bad rate by a little, it becomes much more difficult to identify the causes of this shift.

(Of course, at the end of the scorecard development phase or during the validation phase, analysis should reveal, for a given cutoff, the swap sets in terms of individual characteristics. These are important not only for scorecard variables but also for other variables since they may have a major effect on marketing strategy, staff bonus payments, or even internal politics.)

Implementing with the same acceptance rate is recommended only for a few weeks. In any case, having spent time and money building a new scorecard, one will be keen to begin to accrue the benefits of the development. (Certainly, maintaining the same acceptance rate will usually accrue benefits in terms of a reduced bad rate. However, this benefit will really be seen and be proved only many months later. In many organizations, pressure will be brought to bear to increase the acceptance rate.)

If we can ignore that a new scorecard will usually be replacing an old scorecard, then one approach to setting a cutoff is by assessment of the breakeven point. Let us assume for the moment that we have perfect information on the future performance of groups of accounts in terms of repayment, early settlement, arrears, ultimate loss, and the timing of these events. In such a case, we should be able to assess how much money this group of applications will make for the organization if we choose to accept them. We can produce these data on income and loss for each possible score—or for each score that is a candidate for the cutoff. A simplistic—but not necessarily incorrect—view is to accept all applications that will generate a profit, however small. One could also argue that we could accept those applications whose profit will be zero. This implicitly at least assumes that cases above this point will also generate a profit, although with some products—for example, a credit card—cases that score very highly and are unlikely to generate bad debts also may be unlikely to generate much income.

As was stated, this is simplistic but not necessarily incorrect, at least as a starting point. This falls down mainly because of accounting reasons.

Rather than consider profit, we certainly need to express income and losses in terms of their NPV. This equates future income and expenditure to what they are worth today. For example, is £1000 today worth more or less than £1200 in three years, or a monthly income of £32 for the next three years? The NPV tries to take into account such factors as the interest

one could earn with the money if received earlier as well as the possible opportunity cost of having money tied up for a period.

Rather than consider profit or net income or NPV, we should consider the return on our investment. In lending someone, say, £5000, we are making an investment in that person and in their future ability and likelihood to repay the loan with interest. However, there is a cost in raising the £5000 to be able to lend it. We need to consider how well we have managed the money. For example, if we can lend £10,000 and make a £1000 profit and also lend £5000 and make a £600 profit, in a very simple form, the former represents a 10% return and the latter a 12% return. Clearly, if we could have a portfolio of £1 million of either the former or the latter type of loan, the latter is preferable. (This evaluation of return is not intended to be comprehensive. Indeed, it is extremely simplistic. Many finance and accounting books deal with the matter in much more detail and depth.)

What is important from the point of view of scoring is that we may wish to set a cutoff using return rather than profit. Therefore, we would set our cutoff at a point where all applications will meet some required minimum return threshold. Again, we may be making the assumption that cases with a score higher than this threshold will also meet the minimum return threshold.

In assessing both profit and return, we may run into another accounting issue, the allocation of fixed costs. For example, suppose that our fixed costs are £11 million per annum. These fixed costs may be for the building, staff, marketing, etc. With a scorecard cutoff of, say, 245, we expect to grant 50,000 loans. The fixed costs equate to £220 per loan. However, at a cutoff of, say, 235, we expect to grant 55,000 loans, equating to a fixed cost allocation of £200 per loan. If at a scorecard cutoff of 245 we meet our required profit or return, when we consider whether to lower the cutoff to 235, do we reallocate the fixed costs? Alternatively, do we assume that they have all been taken care of and that we can consider the marginal business from 235 to 244 using only the variable costs in the profit-return calculation? There are complex variations of this, and there is no single answer. Rather, the answer usually rests within the organization's accounting procedures and business objectives.

Another challenge that arises in assessing the profit or return for an application if granted is that, once again, we need to assume that the future will be like the past. We cannot know if the loan will be settled early or prepaid. We can only make some assumptions based on past performance of similar loans. We also have to make some assumptions about the recovery performance of future problem cases—indeed, of problem cases arising from applications that we have not yet even approved.

Therefore, in assessing which cutoff to use, several interrelated credit and accounting issues may need to be considered.

From the scorecard development, we should be able to construct a graph similar to Figure 8.3. For example, at an acceptance rate of only 50%, we have a bad rate of 2%. If we accept 70% of applications, the bad rate climbs to almost 4%. At an acceptance rate of 90%, the bad rate is approximately 7.5%. The triangle represents the current position before implementation of the new scorecard. Clearly, we can take up a position to the right of this, thereby maintaining the bad rate but increasing the acceptance rate. We can also take up a position below it, maintaining our acceptance rate but reducing the bad rate. In fact, any combination above or on the curve and to the right or below (or both) of the triangle represents an improvement on current practice. However, common sense dictates that the only positions worth considering are those actually on the curve.

Often this data is given in the form of a table reporting the results of applying the scorecard to the holdout sample. This table, the run-book, gives the likely effect of each

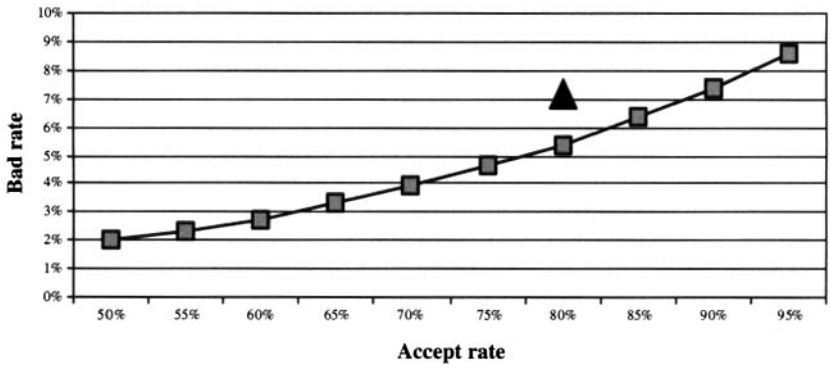


Figure 8.3. *Strategy curve.*

cutoff score on the whole population. Table 8.7 shows a typical run-book with a holdout of 10,000 (9000 goods and 1000 bads). We could also construct alternative graphs that look at score versus profit or score versus return.

Table 8.7. *Example of a run-book.*

Score	Cum. goods above score	Cum. bad above score	Cum. total	Percent of total	Bad rate percent of acceptance	Marginal good:bad odds
400	2700	100	2800	28	3.6	—
380	3300	130	3420	34.3	3.8	20:1
360	3800	160	3960	39.6	4.0	16.7:1
340	4350	195	4545	45.4	4.3	15.7:1
320	4850	230	5080	50.8	4.5	14.3:1
300	5400	270	5670	56.7	4.5	13.7:1
280	5900	310	6210	62.1	5.0	12.5:1
260	6500	360	6860	68.6	5.2	12:1
240	7100	420	7520	75.2	5.6	10:1
220	7800	500	8300	83.0	6.0	8.7:1
200	8600	600	9200	92	6.5	8:1
180	8900	800	9700	97	8.2	15:1

While profit or return is often the driving force behind a scorecard cutoff decision, there are operational factors to be considered. For example, if we greatly increase our acceptance rate, do we have sufficient funds to lend the increased sums of money? Also, have we the operational capacity to deal with more cases being fulfilled and funds being drawn down, as well as the fact that there will be more live accounts and so more customer queries and more redemptions, either early on or at the end of the scheduled term?

If we implement a behavioral scorecard, similar issues arise. Are we about to greatly increase the number of telephone calls or letters that will be required, and do we have the capacity? On the financial side, are we about to greatly increase our credit limits, and do we need to get approval for this, or are we about to increase our provisions for arrears cases?

8.12 Aligning and recalibrating scorecards

If one has a suite of scorecards each built on a separate subpopulation, then it would seem sensible to align them so that a score has the same meaning on all the scorecards. What

this means is that at each score the marginal good:bad odds should be the same on all the scorecards built on the different subpopulations. This is the only part of a run-book that is specific to a scorecard since the acceptance rate and the bad rate are needed for the whole population and their values on individual subpopulations are of little interest. One way to do the recalibration is to run a logistic regression of good-bad outcome against score for each scorecard separately. For scorecard i, this leads to the equation

$$\log\left(\frac{p_G^i(s)}{p_B^i(s)}\right) = a_i + b_i s, \tag{8.19}$$

where s is the score, $p_G(s)$ and $p_B(s)$ are the probabilities of an applicant with score s being a good and bad, respectively, and a_i and b_i are the coefficients of the regression. If this logistic regression is a good fit, one can recalibrate the scorecard for population i by multiplying the score by $\frac{b_1}{b_i}$ and adding $\frac{a_i - a_1}{b_i}$ to all the attributes of the scorecard. This guarantees that all the scorecards have good:bad odds satisfying $\frac{p_G}{p_B} = e^{a_i + b_i s}$. Even if the logistic regression does not fit very well, one should multiply all the scores in the scorecard by $\frac{b_i}{b_1}$ and then use a new variable (to which subpopulation an applicant belongs) in a regression to find the best constants to add to the scorecard to recalibrate it.

What this has done is to align scorecards belonging to different subpopulations in terms of their marginal good:bad odds so that at each score the good:bad odds is a known constant. This idea of recalibrating the scorecard so that it has certain properties can be applied to individual scorecards as well as suites of scorecards. The properties one might want in a scorecard are as follows:

(a) The total score is positive.

(b) The score for each attribute is positive.

(c) There are reference scores that have specific marginal good:bad odds.

(d) Differences in scores have the same meaning at all points on the scale.

However, recalibrating a scorecard to satisfy some of these properties makes it lose others. For example, we have a scorecard where all the attribute scores (and hence the total score) are positive and we want it to have marginal good:bad odds of o_1 at s_1 and o_2 at s_2. If s_1^* and s_2^* are the scores on the existing scorecard that have these marginal odds, then the linear transformation

$$s^* = \frac{s_2^* - s_1^*}{s_2 - s_1} s + \frac{s_1^* s_2 - s_2^* s_1}{s_2 - s_1} \tag{8.20}$$

ensures that the reference scores s_1 and s_2 have the required marginal odds. Since this is a linear transformation, it can be applied to each of the attribute scores in turn, and the result is that the total score is transformed in the correct way. However, it may turn positive attribute scores into negative ones and even introduce negative total scores. Trying to correct these problems will then destroy the two reference score properties.

Thus given an existing scorecard, it is not possible to ensure that it has all the properties (a)–(d). Condition (d) is very like condition (c) in the following way: Suppose that one wanted an increase in the score of s to double the good:bad odds; then this is equivalent to requiring the odds at s_0 to be o_0, at $s_0 + s$ to be $2o_0$, and at $s_0 + ks$ to be $2^k o_0$ for $k = 2, 3, \ldots$. One way to seek a new scorecard that approximately satisfies all these properties and also

keeps the relative rankings in the sample used to build the original scorecard as unchanged as possible is to use linear programming.

Assume that there are N accounts in the sample and they are ordered in increasing original score. Let there be p characteristics and x_{ij}, $i = 1, 2, \ldots, N$, $j = 1, 2, \ldots, p$, are the values of the characteristics given by the ith consumer in the sample. Assume that $x_{ij} \geq 0$ for all i, j. Then the requirement that the ordering is maintained under a new scorecard with weights $w_1, \ldots, w_p$ is that we seek to minimize e_{ir} for $i, r, i < r$, where

$$w_0 + \sum_j w_i w_{ij} < w_0 + \sum_j w_j w_{rj} + e_{ir}. \tag{8.21}$$

This is trying to ensure that the score for the ith consumer is below that for the rth consumer. This has $\frac{N(N-1)}{2}$ conditions, which for reasonable-size data sets make it too large a problem to deal with. We can approximate this requirement by asking that it hold only for nearest neighbors, i.e., that the score for the ith consumer is below that for the $(i+1)$st consumer. This can then be incorporated into a linear program, which also seeks to satisfy the other constraints, namely,

$$\text{Minimize} \quad a_1 \sum_{i=1}^{N-1} e_i + a_2 \sum_{v=3}^{m} (d_v^+ + d_v^-) + a_3 \sum_{l=LM}^{UM} (f_l^{d_+} + f_l^{d_-})$$

$$\text{subject to} \quad w_0 + \sum_j w_j x_{ij} \leq w_0 + \sum_j w_j x_{i+1j} + e_i \quad \text{for } 1 \leq i < N,$$

$$w_0 + \sum_{j=1}^{p} w_j x_{i_k j} = s_k \quad \text{for } k = 1, 2$$

(i_1, i_2 accounts have properties required at scores s_1, s_2),

$$w_0 + \sum_{j=1}^{p} w_j x_{i_k j} = s_k + d_k^+ - d_k^- \quad \text{for } k = 3, \ldots, R,$$

($i_k, k = 3, \ldots, R$, accounts have properties required at scores s_k),

$$w_0 + \sum_{j=1}^{p} (w_j) x_{i_t j} = s_0 + ts + f_t^{d_+} - f_t^{d_-} \quad \text{for } 0 \leq t \leq T$$

(where i_t is the account at whose score the marginal odds are $2^t o_0$),

$$e_i \geq 0 \quad \text{for } 1 \leq i < N, \qquad w_j \geq 0 \quad \text{for } 1 \leq j \leq p$$

(attribute score is positive),

$$d_k^+, d_k^- \geq 0 \quad \text{for } 3 \leq k \leq R, \qquad f_t^{d_+}, f_t^{d_-} \geq 0 \quad \text{for } 0 \leq t \leq T. \tag{8.22}$$

For a fuller description of how this linear program can recalibrate the scorecard, see Thomas, Banasik, and Crook (2001). It is clear, however, that there are psychological and operational reasons to adjust a scorecard so that the scores tend to be in the range 100 to 1000.

This page intentionally left blank

Chapter 9

Implementation and Areas of Application

9.1 Introduction

This chapter deals with some of the issues involved in implementing a scorecard, either an application scorecard or a behavioral scorecard. It also discusses some related issues. In much of the chapter, there is little difference between application and behavioral scorecards; where this difference is significant and is not obvious it is indicated.

The chapter also moves on to deal with monitoring and tracking, defining these separately rather than treating them as interchangeable terms as is often the case. We also look at setting and testing strategies and then try to bring some of the matters together in answering questions such as, "When is my scorecard too old and in need of redevelopment?"

9.2 Implementing a scorecard

For whatever reason, an application scorecard has been commissioned and built, but how do you implement it successfully? To do this, several questions need to be considered.

- Do we have access to all the scorecard elements in the new scorecard?

In theory, one would assume that the answer to this is yes. After all, if we did not have access to the information, how could the data element have been included in the model building? However, several possibilities can arise. For example, it is possible that a data element was collected at the time of the development sample but has since been removed or dropped. In another example, the data elements in some questions may have been collected as part of a retro (i.e., a retrospective analysis) at a credit bureau but are not part of a standard return from the credit bureau.

- Are there any definitional questions as a result of the elements in the scorecard?

One needs to be clear about how the customer or operator will interpret each question. For example, are the questions and answers as used in the development the same as those used now? If we ask about the applicant's marital status, common options are single, married, widowed, divorced, and separated. However, people may consider themselves both married and separated. Other people may consider themselves to be both single and divorced. Whether applicants record themselves as married or separated or single or divorced may change according to a variety of influences, including the region of the country. What is

recorded now may be subtly different from what was recorded in the development sample. The difference may be less subtle, caused perhaps by a change in the order or layout of the options. A member of the staff involved in the application—in a branch- or telephone-based environment—may have an influence on what is actually recorded, by answering the applicant's queries or by the way the questions are phrased. To reduce this effect, a clear policy relating to a wide range of possible questions is needed, and those colleagues who deal with customers should be trained carefully, thoroughly, and repeatedly and their performance continually monitored.

Another area that should be checked is whether the data received from the bureau has changed. For example, in the U.K., from October 2001 some changes were implemented regarding what data are available from the bureau. These changes may lead to fewer searches being recorded against individuals. Therefore, a scorecard characteristic looking at the number of searches in the last six months may change its value.

- What programming is required?

This may seem a trivial point, but even with the advanced state of computer systems used for processing applications and accounts, this often needs consideration. Clearly, if the programming of the scorecard requires the support of a systems department, then the size and complexity of the task may have a bearing on the time scale for its full implementation. One also needs to be very careful with how the implementation and scoring is actually tested. The alternative implementation route is to utilize one of the many pieces of software that are commercially available or, in some organizations, that have been built on to the standard systems. In these cases, the scorecard is parameter driven and can usually be programmed by a user or scoring analyst. However, one should never underestimate the effort required to key in the characteristics and their relative weights and to test that this has been completed successfully. One also should never take shortcuts, for example, to achieve implementation a few days earlier.

- How are we going to treat "pipeline" cases?

In many environments, there will be cases that are in the pipeline at the point we implement the new scorecard. This will generally happen when we have processed an application using the old scorecard but have not yet taken the application to approval or the funds have not yet been drawn down. One reason for this is that we are running a telephone- or Internet-based operation and we have issued a credit agreement for the applicant to sign and complete. Another possibility is where we have approved a mortgage but the property has not yet changed hands and so the funds have not been drawn. In many of these cases, we may be able to adopt the original decision. (In some of these cases, we are in a very weak position if we make someone an offer and then rescind it without their circumstances changing.) However, if, for example, we needed to reprocess the application, we may find that, because of the change in scorecard, the application now fails. Clearly, we need to adopt a policy. One option is to go with the original decision and live with a few marginal cases being accepted for the first few weeks. (This may require some systems solution to override a case that fails and to force it through the system.) Another option is to adhere to the new scorecard rigidly and reverse an original decision, communicating this to the applicant.

9.3 Monitoring a scorecard

In most scoring operations, analysts and managers refer to *monitoring* and *tracking*. In many cases, these terms are used either generically or interchangeably. In this text, we treat these as two separate functions, each with their separate but linked purpose.

Monitoring is considered to be passive and static. It is analogous to the traffic census taker who sits beside the road and logs the numbers of different types of vehicle passing a point in a period of time.

When carrying out monitoring, we should remember that a scorecard provides a prediction of risk for individual accounts but is also used to manage portfolios.

Therefore, some key questions can and should be addressed, including the following:

- Are applications being scored properly?
- Is the profile of current applicants the same as previously, e.g., last quarter, and the development sample?
- Is the current acceptance rate the same as previously and the development sample?

The first question should be required less often than it used to be. Earlier in this chapter, we dealt with ensuring that the scorecard was accurately implemented into our systems. Here we are referring to that area of the application that requires human input, whether by a member of the staff or by the applicant. One common area where some scrutiny might yield dividends is where there is a list of options. This might relate to the classification of occupation, for example. Suppose that the development process used 12 categories of occupation, the 12th one being "other." Suppose also that the percentage of cases scored as other was 5%. We now find that in a current batch of applications, there are 12% scored as other. We need to examine this difference and understand how this change might have happened. There are several possible explanations:

- The 12% is correct and was the comparative figure for the development sample. However, in the development process, some additional effort was spent in trying to allocate these cases of other to one of the other 11 categories.

If this is what has happened, then we could argue that the additional effort was wasted. Our general assumption of the future being like the past has been invalidated by our extra investigation.

- The operators who are classifying occupation are not being as effective as they should and, if they do not immediately know the category into which an employment should fall, they classify it as other and move on.

In this case, the issue becomes one of training and perhaps motivation. We have seen this type of thing happen in one organization split into three regions, where the percentage of cases classified as other in the three regions was 8%, 12%, and 26%. The major reason for these differences was the different emphases on data quality and process efficiency placed by different regional directors.

- There has been a change in our applicant profile.

If this is true, then we need to investigate the cause of it, what can be done about it, and the effect on the scorecard. One possible cause is that new occupations are developing all the time and our processes do not always keep pace with these. For example, 5 or 10 years ago, there were far fewer website developers than there are now. Some further comments on this area can be found in (Edelman 1988).

Another possible cause of a change in our applicant profile is that we have seen a change in the type of person who is applying. This may be caused by a change in the

organization's marketing strategy, which may employ mailing, television advertising, or press campaigns—or a mix of these. Alternatively, the organization could have changed policy or strategy on a whole subset of the potential applicant population. For example, there could have been a marketing drive to attract students or a loan promotion aimed at gold card customers. Despite many years of recognizing the mutual advantages of communication between credit and marketing strategists, it still happens too infrequently and to too little effect. Most managers of scoring operations will be able to recount tales of discovering changes in marketing strategies after the event, by which time they have declined a high percentage of the additional business that was attracted. Worse is that if a poor-quality profile of applications is generated by a campaign, even after the high decline rate, the average quality of the accepted propositions will be poor. Therefore, while some of the marketing department may be content with a high response, most of them will be unhappy that the credit department declined many of them. Of course, the credit department is also unhappy because the quality of the book of accepted cases is being diluted by a poor-quality batch of accounts.

A further possibility is that there has been a change in the marketing strategy of a competitor. This can happen if the competitor is relatively large or where their offering or positioning is similar.

A change in profile may be caused by a change in the economic environment and outlook. If the economy suffers a recession or an expansion or if forecasts are that either of these will occur, we may see changes not only in the numbers of applicants for credit but also in the quality.

Having discussed at length some of the causes of a change in our applicant profile, let us move on to consider the measurement and the possible resolutions.

Table 9.1 is a common type of report. The characteristic analysis report takes each attribute of a characteristic and considers the differences in the proportions with each attribute between the development sample and a current sample. It then calculates the effect on the score for the characteristic.

Table 9.1. *Characteristic analysis report.*

Attribute	Development sample percentage	Current sample percentage	Score	Difference percentage	Difference percentage x score
Employed full time	52	42	37	–10	–370
Employed part time	9	16	18	7	126
Self-employed	18	23	15	5	75
Retired	11	10	28	–1	–28
Houseperson	6	4	11	–2	–22
Unemployed	1	3	3	2	6
Student	3	2	8	–1	–8
					–221

Thus for this characteristic, the average score has fallen by 2.21 points. We may find other characteristics where the change in average score is greater or less than this, and it is the total of the changes in the scorecard characteristic that determines the total change in score between the development sample and the current sample.

Therefore, we should do this type of analysis in three ways. The first is to examine each of the characteristics in the scorecard. The second is to do a similar analysis for the total score,

and the third is to look for significant changes in nonscorecard characteristics. Significant changes here may reveal a significant change in our applicant profile or a characteristic that should be included in the scorecard, especially if it is highly correlated with performance. Clearly, we cannot construct a meaningful characteristic analysis report for a characteristic not in the scorecard (as the score weights will all be zero). However, we can construct a population stability report, as in Table 9.2.

Table 9.2. *Population stability report.*

Score	Development sample percentage	Current sample percentage	$B-A$	$\frac{B}{A}$	$\ln\left(\frac{B}{A}\right)$	$C \times D$
	A	B	C		D	
<200	27	29	0.02	1.074074	0.07146	0.0014
200–219	20	22	0.02	1.1	0.09531	0.0019
220–239	17	14	–0.03	0.823529	–0.19416	0.0058
240–259	12	9	–0.03	0.75	–0.28768	0.0086
260–279	10	11	0.01	1.1	0.09531	0.0010
280–299	8	6	–0.02	0.75	–0.28768	0.0058
300+	6	9	0.03	1.5	0.40547	0.0122
					Stability index =	0.0367

We need to consider how to interpret this.

Statistically speaking, the stability index is an X^2-type of measure, and therefore the interpretation of the stability index should incorporate some indexing. We may wish to consider using degrees of freedom, especially because the number of cells used may affect the value of the index.

Some have suggested a rule of thumb to the extent that a stability index of less than 0.1 indicates that the current population is fairly similar to that from the development sample. An index of between 0.1 and 0.25 suggests that we should pay attention to the characteristic analysis reports to see if we can identify shifts within the characteristics. However, an index value above 0.25 indicates some significant changes in the score distribution.

In general, however, this is a tool that is fairly easy to use but lacks some sophistication and consistency. Other measures could be used that would counteract these objections. To measure the significance of any difference between the score distributions at the time of the development and the current distributions, one could calculate a Kolmogorov–Smirnov test statistic and test its significance. Alternatively, one could calculate the Gini coefficient.

However, of utmost importance is that, on a regular basis, one goes through the process of measuring the stability of the characteristics and the score distribution. Even if one has no notion of significance, if the index or measure is growing, this alone is a sign that there is some population drift.

We need to consider one other point here. When developing a scorecard, we can and do use exactly the same measures although then we are looking for large significant differences to support the development and implementation of a new scorecard as being better than the old one. Once we have implemented the new scorecard, we are looking for insignificant values to suggest that there has been little change in applicant profile and so forth since the development.

9.4 Tracking a scorecard

While monitoring is passive and static, tracking can be considered to be active and dynamic. If we consider the two analogies mentioned above, tracking is more like following an animal to see where it has its young or like a police car following another car to assess its speed or quality of driving.

Tracking principally allows us to assess whether the scoring predictions are coming true. Following that, we need to consider whether the predictions are coming true for realistic subsets of the population. Also, if the predictions are not coming true, we need to consider what action to take. When tracking, as with monitoring, we should remember that a scorecard provides a prediction of risk. Therefore, one of the key reports is to examine how accurate the scorecard's predictions are. Thus, the key questions are the following:

- Is the scorecard ranking risk properly?
- Is the portfolio bad rate what we would expect given the development sample or the distribution of accounts booked?
- What changes can we detect between different subsets of the account population, defined by segments, or cohorts, or by different account strategies?

To assess if the scorecard is ranking risk properly, we need to examine whether at each score the percentage of cases expected is approximately what happens. In Figure 9.1, we plotted our expectations from our development sample along with data from three different samples.

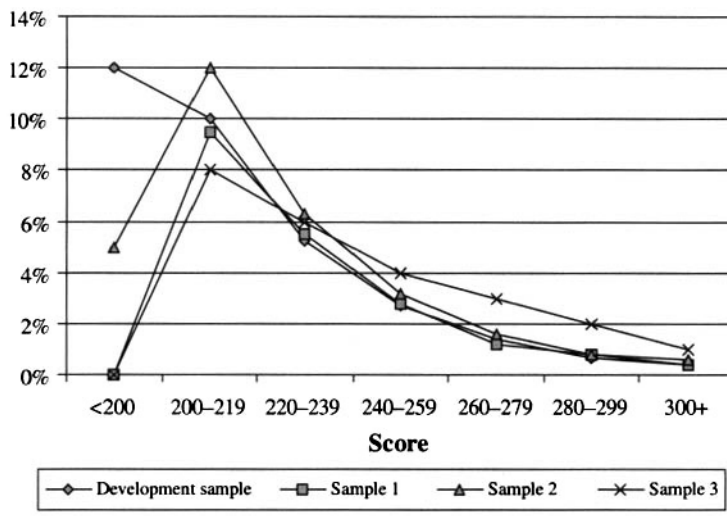


Figure 9.1. *Bad rates.*

Clearly, we are operating this portfolio with a cutoff of about 200. With the first sample, our bad rates for each scoreband are approximately what the scorecard predicted. There are some fluctuations around the development sample, but on the whole it is well aligned.

The second sample conveys some different messages. First, note that in this sample we took on some business below our cutoff. The bad rate in this region is below that of the first region above cutoff, so we might conclude that there has been some intelligent overriding in

this area. Above the cutoff, the bad rates are consistently above expectations. This suggests that the economic environment in which this sample is operating is worse than that in which the development sample operated. Note, however, that the scorecard is still working well since the curve has a similar shape. On the other hand, if the second sample comes from a different type of business, e.g., newspaper advertising rather than branch-based applications, then we need to consider whether we should amend our process. Another possibility is that we raise the cutoff so that the profit or return at the cutoff is the same. Of course, if the applications from newspaper advertising are cheaper to acquire, it may be that we can accommodate a slightly higher bad rate at each score and still produce the same return.

With the third sample, we see lower bad rates than expected just above the cutoff and higher bad rates than expected at higher scores. As this curve is beginning to move away from the curve we expected and tending toward a flat horizontal line, we might suspect that the scorecard is not working as well as it should.

We should consider two other points here. First, sample 3 might actually be making more money. For example, a 50% reduction in bad accounts in the 200–219 scoreband and a 50% or even 100% increase in bad accounts in the 260–279 scoreband, say, might mean that overall we have fewer bad accounts. Obviously, one of the key factors in this calculation is the number of accounts in each scoreband, although with careful marketing and targeting, we should be able to exert some control over the number at each scoreband. Nevertheless, the flatter shape of the curve does suggest that the scorecard is not working well on this sample of our business.

The other point worthy of consideration here is that we ought to place greater focus on what is happening at and around the cutoff or at and around candidate cutoffs. If we are operating this portfolio with a cutoff of 200 and find that cases scoring in the 260–279 scoreband have a higher bad rate than was expected, while that may be a nuisance, we are never going to avoid these accounts. First, the actual bad rate for this high scoreband is still likely to be lower than for the scoreband just above cutoff, even if the actual bad rate at that point is below expectations. Second, if an application scores in the 260–279 band, a redeveloped scorecard will most likely still score it well above any feasible cutoff.

If we move away from application scoring to behavioral scoring (or even to collections scoring, attrition scoring, etc.), the principles are very similar. If we can wait for the outcome to appear, which in these cases may be as soon as 6 months later, rather than 18 to 24 months for an application scorecard, we can assess how well the scorecard is performing on different parts of the portfolio.

Now the problem is that if we wait 24 months to discover that we have a weakness, we have taken a lot of business during that 24-month period that we would now rather not have taken on. Therefore, one key objective would be to make some sort of interim assessment.

These interim assessments can be carried out using some simple tables and graphs that are known by a variety of names: dynamic delinquency reports or delinquency trend reports, or vintage analyses or cohort analyses. The fundamental issue is that for each cohort of business, usually one month's business or one quarter's business, perhaps split also by another variable—customer-noncustomer, branch-telephone, score—we compare the performance at similar stages in that cohort's life.

Therefore, suppose we take as a cohort a month's business and record its performance at the end of each subsequent month. We can build up pictures of how this cohort is performing by recording the percentage of accounts that are one down, two down, or three or more down. We can also record the percentage of accounts that by the end of the month in question have ever been one down, two down, or three down. We can do the same analysis by value rather than by number, taking either the loan balance or the arrears balance as a percentage of the

total amount advanced. This could lead to 12 graphs or reports for each subset. However, these graphs or reports encapsulate the whole history of the portfolio from one particular point of view.

An example of this appears in Figure 9.2. In this graph, we plotted the performance of 12 cohorts of business from April to the next March. For the oldest cohort—business from April—we examined its performance at the end of each month from the third month of exposure until the 15th month of exposure.

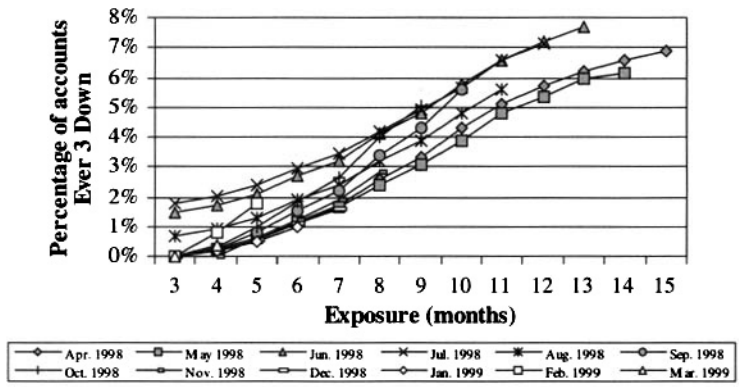


Figure 9.2. *Dynamic delinquency report.*

(One needs to be careful about how a cohort is defined. This may vary from product to product. One could define a cohort by the month or quarter in which the application was made, or the application was approved, or the money was drawn down, or when the account was activated. Further definitions include the period in which there was first financial activity or the period in which the first payment was due. It is obviously important that the definition is applied consistently. It is also important that the definition is consistent with the objectives of the analysis and with the product. For a credit card, one could use the month in which the account was activated, if it requires activation prior to use, or the month of the first debit activity, i.e., transaction, or the month the card was issued. For a loan, the better choices would be the month or quarter the funds were drawn or the month or quarter when the first payment is expected.)

Let us assume that the definition used in Figure 9.2 is from the month in which a loan payment is first due. Therefore, April 1998 business is a tranche of loans that have a first payment due date in April 1998. We have 15 months of exposure for this cohort; i.e., we looked at this group of loans at each month until June 1999.

For the next cohort of business—May 1998—we also looked at its performance every month until June 1999, but that makes only 14 months of exposure, and so on.

This graph looks at the percentage of accounts in each cohort that ever got to being three payments down. This is one of the more common pictures as this is often the scorecard's definition of bad.

We can see a number of things from this graph:

- The graph starts at only the third month of exposure. Accounts cannot miss three payments until three payments have been due.

- All curves are either flat or increasing. Because we are looking at a worst-ever status, the percentage of accounts reaching this stage cannot ever decrease as we lengthen the

exposure. (This relies on our systems retaining a record of accounts even if they are closed.)

- By examining the curves, we can establish a common pattern and then are able to identify deviations from this.

- If a curve is below the general trends, this cohort of business is performing better than the norm. Such a deviation from the norm should be investigated and, if possible and appropriate (taking into consideration profit, etc.), we may wish to attract more of this business. On the other hand, a more radical approach here might be to increase the level of risk by, for example, reducing the level of checking a part of the credit assessment. In other words, bring the business back into line by relaxing our credit assessment process, increasing the risk of each case, but accruing some financial benefit from the relaxation of the credit assessment. For example, if the organization asks the applicant to supply proof of their income, we could increase our risk and reduce our costs by dispensing with this requirement. Another option is to reduce the level of credit reference checking, perhaps not checking on the previous address, irrespective of the length of tenure at the current address.

- If a curve is above the general trend, this cohort of business is performing worse than the norm. Such a deviation should be investigated and, if possible and appropriate, we may wish to try to attract less of this business. If we can identify this very early on, then we can take early corrective action. In Figure 9.2, July 1998 is performing worse. If we identify this through the percentage of cases one payment down and can identify this after three months, i.e., at the end of September, then the change will begin to take effect on some of October's business and certainly November's business. The longer we leave it, the more cohorts of business may also perform poorly. Of course, we must also be careful that we do not act prematurely without proper analysis and consideration of the cases of the poor performance. Alternatively, we could introduce additional steps into the process to improve the credit quality.

- April 1998 and May 1998 business look similar. We then have a worse pattern for June 1998 and July 1998 business. However, these curves are not steeper than the previous two; rather, there is a parallel shift upward. What appears to have happened is that this tranche of business is about 1.5% higher than the norm. A possible reason for this, especially since the shift occurred from the beginning of the plotted history of the performance of the cohort, is that we were hit by some increased fraud. Once this increment is removed, the credit quality of the rest is similar. By August 1998, we appear to have begun to sort this out as the parallel shift is reduced.

- September and October 1998 business is clearly worse than the norm. The curves are steeper, suggesting a poorer quality of business or poorer performance. November 1998 is almost back in line with April and May 1998, suggesting again that we have sorted things out. This could be due to a change in the process, or a change in the scorecard cutoff, or a change in marketing strategy. However, February 1999 again looks to have a worse performance, perhaps caused by the post-Christmas rush for further credit among many consumers.

As was stated earlier, the same type of graph can be constructed for different definitions. We can clearly use Ever One Payment Down and Ever Two Payments Down. We can also use Currently One Payment Down, Currently Two Payments Down, Currently Two or More

Payments Down, etc. In these cases, obviously, the curves can go down as well as up as accounts in arrears can improve from one month to the next. In fact, later on in the life of the cohort, once most of the bad cases have been flushed out, we might expect to see these curves fall as our collections and recoveries activity corrects more accounts each month than the maturity process pushes into arrears.

Clearly, the reports and graphs looking at earlier stages of delinquency can be used to identify at an early stage if a cohort of business appears to be performing worse than the norm. For example, if we see a higher rate of one-down cases after three months and can take corrective action immediately, we have only three months of business taken on with the current process and procedures.

We can also use percentages of values rather than the percentages of numbers. We should recognize that moving into current arrears or value may represent a step away from tracking the performance of the scorecard because scorecards in general work on the percentage of accounts going good and bad, rather than the percentage of the monetary value of the accounts. However, it is a step toward managing the portfolio.

To return to scorecard tracking, we can also produce these graphs for a particular scoreband, thereby controling one of the key factors—the quality of business. Clearly, we can get a better or worse tranche of business by changing our cutoffs or our marketing. To some extent, if we make a conscious decision to do this, we would expect the effect to show up in a dynamic delinquency report. However, if we were to produce a report for, say, scoreband 220–239, then irrespective of the number of accounts that fall into that band in each cohort, we would expect the performance to be similar. Any deviation from the norm could be a sign that the scorecard is deteriorating. Of course, we might also find that for one cohort in this scoreband, the average score is 231, while for another it is 227, so we may need to control further by cutting the the scorecard into even narrower intervals, provided we have enough data to make analyses and conclusions reliable.

We could also produce the same reports and graphs for specific attributes. For example, we might produce one graph for homeowners, or one for customers aged 25–34, or one for customers whose loan is between £5000 and £8000. Obviously, the more we cut the data, the fewer cases there are, leading to more volatile results and less certainty in our conclusions. If we have so little data that we cannot be certain about our conclusions, then we might decide that it is not worthwhile to carry out the analysis.

These graphs allow the user to easily identify trends that happen at the same point in a cohort's exposure. They will not easily allow one to identify trends or events that happen at the same point in real time. For example, if we expect arrears to rise post-Christmas 1998 and then fall in early spring, this will affect each tranche at a different point in its exposure. For April 1998 business, January 1999 is month 10, while for August 1998 business, it is month 6. A much clearer example of this happening occurred in the late 1980s in the U.K., when there was a six-week postal strike. Although cardholders could make their monthly credit card payments through the banking system, many payments were received late, if at all. Cardholders claimed—and some believed—that if they did not receive a monthly bill, they were not due to make a payment. In terms of current arrears, the percentage of cases one payment down on some portfolios rose from, say, 10% to 30%. At the end of the postal strike, the wave had been started and the percentage of cases two payments down rose from 4% to 7%. However, three to four months later, the wave had dissipated and there was minimal effect on ultimate write-offs. (Indeed, any adverse effect on the ultimate write-offs was more than compensated by the additional interest income.)

One may also wish to identify whether there are any seasonal effects. For example, it may be that the February business, because of lower quality, always performs worse.

In some portfolios, the cutoff has been raised for a short period at the start of the year to counteract this.

Once again, the above focusses on application scoring, so how does this type of analysis work for behavioral scoring? To a great extent, the application of the analysis is exactly the same. The score provides a prediction of the percentage of cases of a particular type that will behave in a particular way within a period of time. Therefore, we should track the performance of these accounts after they have been scored and dealt with accordingly to assess the accuracy of the scorecard and the consistency of their behavior. What we might need to be more careful about is the definition of a cohort of business, although factors commonly used in such definitions include time on books and the size of the limit or balance.

We also need to bear in mind that behavioral scoring is often used in conjunction with adaptive control strategies. With these, score may be only one of a set of factors that are used to determine how to manage the account and how to react to the customer's behavior and the account performance. We deal with this more in section 9.6. The key point here is that not all accounts scoring, say, 500–539 will be dealt with identically. Also, the range of actions for them could be quite different from those accounts scoring, say, 580–619 and so we should not necessarily expect a smooth curve, even in theory.

9.5 When are scorecards too old?

Through tracking and monitoring, we can begin to assess when the applicant population has changed significantly from our development sample, in terms of both the demographics and other data available at application stage and in the performance.

There is no simple answer or simple statistical or business test that can be performed to decide when corrective action is required and what it should be. Scorecard performance degrades over time for a number of valid reasons. There is then a business decision to be taken to weigh up the advantages and disadvantages of a new scorecard, bearing in mind the feasibility of building a new one. For example, do we have enough data with enough mature performance and a sufficient number of bad accounts? For example, if our scorecard has degraded because of a change in marketing strategy, will a new scorecard suffer the same fate?

Before we set about redeveloping the scorecard in its entirety, there are a number of possible modifications that will either extend the life of the scorecard or improve its effectiveness through the rest of its life.

The first thing we can do is examine any potential for realignment. Suppose we have a scorecard where we have a characteristic of time with bank, as in Table 9.3. We have been tracking the performance of accounts based on their total score but looking at the different attributes of this (and other characteristics) separately. This produces a graph similar to Figure 9.3.

Table 9.3. *Time with bank.*

Time with bank (years)	
Attribute	**Score**
0	5
<1	7
1–3	18
4–6	31
7–10	38
11+	44

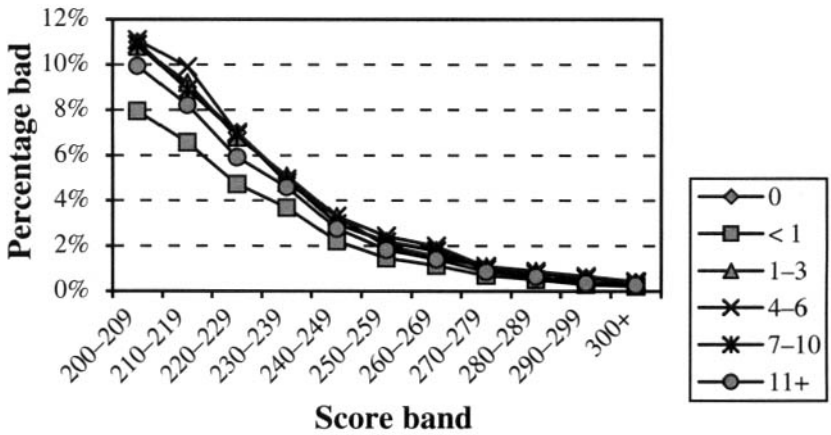


Figure 9.3. *Alignment analysis: Time with bank.*

What we would expect to see if the scorecard is working is that at each scoreband, the percentage of bad cases is the same whatever the time with bank. In other words, the scorecard has adjusted adequately for the different levels of risk represented by the different tenure of the existing banking relationship.

What we see here is that the pattern is similar except for applicants who have been at their current bank for less than one year. What could be causing this persistent anomaly?

A possible reason is that we are more stringent in taking on these applicants as customers, and we may find that we filter out some of them. For example, if we use some other step in the credit assessment—a credit reference search, for example—we may find that for those that score 200–209, say, 88% of them, are accepted. However, for those who have been at their current bank for less than one year, this percentage is only 70%.

Another possibility is that there has been a shift in their level of risk as measured by the scorecard itself. This might be the case here if we find out that banks are being more selective in taking on new customers. Therefore, anyone who has recently changed banks has also recently passed through the new bank's assessment procedures and therefore is better, on average, than the corresponding group in the development sample.

Other reasons could be hypothesized, and some may even be capable of being proved or disproved. However, from the point of view of managing the lending business, all we need to do is to understand that the effect is happening and that it is happening fairly consistently. Once we have established that, we can take steps to deal with it. In this case, applicants with a time with bank of less than one year are performing similarly to other applicants scoring approximately eight points more. Therefore, one simple adjustment is to change the scorecard weight for this attribute from 7 to 15 points. This not only reclassifies all of the existing accounts but also introduces some accounts that previously scored 192–199 and will now score eight points more and achieve the cutoff of 200. We have not seen how these will perform, but with such a smooth pattern, one may be comfortable assuming that they should perform in line with the new score calculated for them.

If there is some concern about taking on new business, this can be introduced on a trial basis, for a month, say, or one could take on a random 1 in 10 of these additional accounts until confirmatory evidence has been gathered. On the other hand, if one does not introduce additional business, then the scorecard realignment becomes a hypothetical and philosophical exercise, i.e., if we are not prepared to act, why bother doing the analysis?

Similarly, if an attribute is underperforming—accounts are performing worse than expected—we can reduce the score weight and, as a result, will now decline cases that previously would have been approved.

(When looking at percentages of bad cases, there are many situations for which we need to consider carefully of what it is that we are taking a percentage. Not only may we wish to look at numbers and financial values separately, but depending on the product, we may have to look separately at accounts opened, accounts still open, active accounts—debit active or credit active or either. We also need to ensure that our computer and accounting systems include accounts written off as bad and carry the actual loss somewhere—although the account balance has been set to zero by the write-off.)

Therefore, it is possible to realign the scorecard, reflecting gradual or developing changes in performance. In general, one should not be too alarmed. It is recognized that scorecards do not suddenly change their performance. Changes happen gradually. However, eventually a decision to redevelop has to be taken. The principle reasons for this are the following:

- Minor realignment either is impossible or has become too complex to understand what is and has been going on.
- Better quality data are now available.
- There are new information sources.
- The goals and practices of the organization or the marketing strategy have changed.

Monitoring and tracking, each in their own way, will provide clues to when the point has been reached that a redevelopment is required. Other factors also need to be considered. For example, the economics of the scorecard and its redevelopment may have a bearing on the timing decision. For example, the redevelopment could take three to nine months (or even longer) to complete and be running live. For example, if the strategy of the organization is likely to change again soon or has changed fairly recently so that there is little performance information available, this may affect the timing decision.

On the other hand, Lewis (1992) puts forward the view that "only rarely are credit scoring systems replaced due to failure to perform as expected" (pp. 115–116). Indeed, he viewed the replacement of the scorecard as being a task of medium-term to long-term planning. "Credit management are well advised to give thought in advance to the degree of degradation of the performance of their scoring system that would cause them to replace it. Since system performance degrades slowly, if at all, if internal policies have not changed, management can predict when the replacement should take place and can plan the assembly of the sample data that will be needed for the development of the new system."

9.6 Champion versus challenger

The use of champion and challenger strategies became much more widespread through the 1990s and is now more accepted through credit and other areas. The principle is quite simple. There is an accepted way of doing something. This is known as the champion. However, there are one or more alternative ways to achieve the same (or a very similar) objective. These are known as the challengers. However, there is no evidence of the effect of the challengers. Therefore, on a random sample of cases, we try the challengers. This trial not only will test effectiveness in comparison with the champion but also will allow us the

opportunity to identify the existence and extent of side effects. Eventually, we decide that one of the challengers is better overall than the champion, and this challenger becomes the new champion.

This occurs in medicine. There are established ways to treat specific ailments. These established ways might be by drugs, counseling, etc. Suppose that we are developing a new toothpaste. We have passed a number of clinical trials in which the chemical composition of the toothpaste was assessed, and we also tried it out on laboratory animals. Eventually, we need to try it out on humans. The key factors in assessing the toothpaste's effectiveness include the taste, how clean the teeth feel, the color of the teeth after brushing, and the ability of the toothpaste to prevent tooth decay. Of these four factors, the first three are fairly subjective, while the fourth takes some time to establish, and we must test the toothpaste on large numbers of humans to establish average effectiveness. (We cannot easily, for example, brush half of each person's teeth with one toothpaste and half with another.) The existing toothpaste is the champion while the new toothpaste is the challenger. In carrying out a trial, we try to capture meaningful figures on the key factors as well as record other measures. For example, was there a change in the incidence of mouth ulcers? Did the subjects endure more headaches or nausea?

Television advertisements abound with similar cases, not necessarily to do with toothpaste. Laundry detergent is a common example. Within medical research, there are other experiments, for example, in cancer treatments. Within the credit industry, we also carry out such experiments. Indeed, a person who has a credit card is almost certain to have formed part of a challenger group at some point without knowing about it.

In the credit environment, we often refer to the whole process as adaptive control. By recording and analyzing customer reactions to our management and control actions, i.e., experiments, we can learn how best to control their behavior to maximize some objectives, such as profit and retention. One might expect a behavioral psychologist to use similar approaches and terminology in controlling animal behavior.

We can carry out such experiments at the application stage. For example, for loan applicants, we could assess their affordability and, on a random sample, select those with higher affordability and offer them either an increased loan or a credit card as well. The desired benefit would be increased income. However, it might be that those who are more susceptible to such an offer are also those who are more likely to end up in trouble, and so our increased bad debt might wipe out the increased income.

Another example is in a telephone-based lending operation, where, after the customer has received an agreement in principle, the customer is sent an agreement to sign and is asked to return this, completed, along with some documentary proof of income or employment or residency. Not everyone completes this process. Some people do not because they change their minds about borrowing the money or about the planned purchase. Others do not because they find another lender with less-stringent requirements for proofs. A challenger strategy is to remove the requirement for the documentary proof. This can be trialled for a short period to see the effect on the percentage who complete the process. Of course, removing the requirement for proof will increase the risk of fraud, and so this will have to be monitored and taken into consideration when assessing the effectiveness of the challenger.

Another challenger strategy would be to offer the customer an incentive to complete the loan—i.e., complete the loan documentation so that the loan can be drawn—or to complete it within a specific period. Here, running strategies in parallel will allow us to measure the relative cost-effectiveness of offering different incentives.

There are other potential impacts from such a challenger. It may be that if customers who are asked to send in proof of their income receive their agreement in principle a few

days before they are due to be paid, they wait until they have an up-to-date payslip available. If this proof is not required, they may return their agreement, not only in greater numbers, but by return of post. Therefore, it is possible that the flow of returned agreements would be affected.

Probably the most common use of champion and challenger strategies within credit is in the area of managing accounts. There are several areas in which we can operate experiments, and the systems for managing accounts often differentiate between these. Trivially, in most cases, when we talk about account management, we are talking about accounts with something to manage. These are, therefore, revolving credit accounts. They may be overdrafts, or current accounts, or flexible mortgages, but the most common area is credit cards. Within credit cards, there are at least six areas with which account management deals. These are as follows:

- Overlimit—how to deal with accounts where the balance is over their agreed credit limit.
- Delinquent—how to deal with accounts that have missed one or more due payments. These may include accounts that are also overlimit.
- Credit limit—how to manage the cardholder's agreed credit limits.
- Authorization—how to deal with requests for authorization from merchants. These will be generated either because the proposed purchase is above the merchant's limit or because the proposed purchase will take the customer above some threshold, e.g., their credit limit, or a system warning on the number of transactions in a period.
- Reissue—whether to reissue a card to a cardholder at the expiration date of the card (or when it is reported as lost) and, if so, for what period.
- Marketing—how to maximize retention or cross-selling opportunities.

As an example, we can consider how we deal with cases that miss a payment and become one month delinquent. The credit card organization will have an established way to deal with these. Years ago, this might simply have been to send the cardholders a letter reminding them of their responsibility. However, we can construct a challenger to this strategy. We have a range of options open to us:

- We can send the customer a letter. Indeed, we can have different strategies using a range of letters varying in the severity of the tone and what we are threatening to do in terms of further action.
- We can telephone the customer, either at work or at home.
- We can send someone to the customer's home to talk.
- We can generate a message on the statement.
- We can ignore the missed payment.

How do we decide which action to take for which customer? Well, this decision would depend on a range of other factors, including the following:

- The arrears history. If this is the first time this has happened, we might wish to take a firm but friendly stance. On the other hand, if the customer regularly misses a payment and then makes it up the next month, it might be worth ignoring it and seeing if the customer makes the payment. Similarly, the time since their last delinquency (and the severity of it) might have a bearing.

- The value of the balance or the arrears. In terms of cost-effectiveness, action that might be appropriate for a £4000 balance where there is a £200 payment that has been missed might not be appropriate for a £200 balance where the missed payment was for £5.

- The relative value of the balance compared with their normal usage patterns, following on from the above. If the customer normally has a £1000 month-end balance, but this month, when the payment was missed, the month-end balance has risen to £4000, this could be a sign that there is an increased risk, and so appropriate, firmer action is required. On the other hand, it could also be a sign that the customer is away from home—on holiday, perhaps—and is using their card, having forgotten to arrange for a payment to be made.

- The behavioral score, which represents the probability that an account in a certain current state will reach some other state at some future point. For example, for accounts that are up to date or two or fewer payments down, the behavioral score might predict the probability that the account will reach three payments down within the next 6 (or 12) months. Therefore, even for accounts that are one payment down or are even up to date, if the customer has a low behavioral score, we ought to treat them differently from those with a high behavioral score.

- How long one has been a customer. This is similar to consideration of arrears history. We may wish to deal with relatively new customers in a different way from more established customers.

- Other relationships. We may wish to take into consideration other accounts held by the customer.

- The overlimit status. If the customer is also overlimit, this will probably call for a more stringent approach.

- The card is due for reissue. If the card is due for reissue, we may be able to use the promise of reissue to encourage the cardholder to get their account back in order. Alternatively, we can take the opposite view and decide not to reissue the card.

Account management packages allow the user to select from a large number of factors in setting a strategy. Each strategy, therefore, can have several nodes. What might happen in a collections area is that there is a champion strategy that operates on 80% of the accounts and that we run with two challengers on 10% each, one that is more stringent and one that is less stringent. As with the applications strategies (and the medical experiments), there are several factors to be assessed with the three strategies. However, usually the overriding consideration is of profit or return. Therefore, we need to understand the cost-effectiveness for each strategy. To do this, we need to consider several areas:

- The cost—of letters, telephone calls, visits.
- The effectiveness: What ultimately happens to the accounts? Do the accounts get back in line? Do we end up writing off bad debts? Do customers get annoyed at our approaches and take their business elsewhere?
- The positioning: Have we generated customer complaints? Has there been an effect on our public position?

We can then run these strategies for a few months and monitor the effect on many measures. We need to monitor income, and also how many cases pass on to the next stages of delinquency. We need to record customer complaints (and even customer praise). Costs will need to be allocated quite carefully. However, after a few months, we should be able to assess if any of the strategies is coming out on top or at the bottom. Clearly, if one is coming out poorly, we remove it and either move back to using the champion on that group of accounts or develop a new strategy.

If one of the challengers is coming out on top, we should expand the size of the group of accounts being managed using this strategy. However, this requires that we consider carefully the resource implications. For example, the best challenger may involve making telephone calls instead of sending out letters. We may have reached our capacity of outbound calls until we recruit more staff. Alternatively, the number of telephone lines the system will support might restrict us. Also, when making outbound calls, if we are unable to contact the cardholder and leave a message, we may greatly increase the number of inbound calls, with the obvious resource implications.

Running these competing strategies and allocating cases randomly and being able to monitor the outcomes and allocate costs to assess the effectiveness places great strains on our computer systems. Account management packages provide support in these areas. However, they are quite expensive and so are probably of use only when the portfolio is of significant size. How large it has to be to justify the cost will obviously depend on how much additional profit can be squeezed out of the accounts by better management. In addition, there is the view that if we can manage existing accounts much better and maximize our profit from each one, then we can take on much higher risks at the front end. Therefore, a good strategy management system, together with good strategy managers, will not only increase profit from the accounts but will allow more applicants to be approved.

Systems will also allow the user to randomly select accounts. To do so, most systems have a random number allocated to each account. This might be a two- or three-digit number. However, the best systems will have more than one random number allocated. This is useful in reallocating test groups.

In the example above, if one challenger proves to be ineffective, we really want to randomly reassign these cases to the other strategies. If we replace the ineffective strategy with a new strategy, then we are not, strictly speaking, testing the new strategy. Rather, we are testing a combined process of using the now-discarded strategy followed by the new strategy—in statistical terms, there may be a carry-over effect—and accounts managed in this way may not perform in the same way as accounts managed by the new strategy alone. It should also be recognized that, if the strategies are significantly different, this may generate customer confusion and, perhaps, complaints.

This page intentionally left blank

Chapter 10

Applications of Scoring in Other Areas of Lending

10.1 Introduction

In the previous chapter, we dealt with the most common applications of scoring or of a scoring approach in lending. In this chapter, we look at some other areas of lending where a scoring-type approach can be used. In Chapter 11, we look at some other areas where a scoring-type approach can or has been used but which fall beyond the normal confines of a lending environment.

The topics we touch on in this chapter are addressed in order of the process, but not necessarily in the order of their priority or importance. These topics are prescreening, preapproval, fraud prevention, mortgage scoring, small-business scoring, risk pricing, credit extension, transaction authorization, aspects of debt recovery, and provisioning for bad debt.

In almost all of these cases, there is a lending decision to be made and there are either two possible outcomes or we can reconstruct the situation so that it looks as if there are two possible outcomes. In a few cases, there are clearly more than two possible outcomes, and we try to identify these and comment as appropriate.

10.2 Prescreening

In principle, prescreening is where we make some or all of the lending and associated decisions before the applicant is aware of what is happening. In some cases the applicant will never be aware of what has happened. For example, before sending out a mailing, we may be able to match the name and address against a file of previous bad debtors and remove those who have a poor credit history. Clearly, in such a case, those who receive the mailing are not explicitly aware that they have passed through some form of initial test. Of course, those who are excluded from the mailing are certainly not aware that they were initially considered and then filtered out. In this example, the file of previous bad debtors may be an internally held file of the institution's previous borrowing customers. Alternatively, it could be a list from a credit bureau. (However, the reader should be aware that different restrictions exist in different countries regarding exactly how this information is used.)

Before approaching the customer with an offer, we can carry out a wide range of prescreens. Perhaps we merely wish to match or confirm the address. On the other hand, we can carry out several prescreen tests and, if we have enough information, we can even carry out a full assessment and maybe make a firm offer (preapproval). Another option open to

us is that, on the basis of the results of the prescreen, we may elect to make the customer a different offer.

When considering using prescreening, one needs to carefully consider the cost and benefits. Using third-party information comes at a cost. Also, if using previous serious arrears as a prescreen rule, it may be that the population one is about to approach has few members who would be removed from the list. Therefore, the information is useful in theory but not on the particular target population. It may also be that the credit-assessment process would identify those higher-risk cases that respond. Therefore, one needs to consider the trade-off between removing a large number of high-risk cases from the mailing list and trying to identify the smaller number of high-risk cases among those who respond.

One other point may need to be borne in mind. In building a traditional scorecard, we carry out reject inference on those who applied but either were not approved or chose not to take out a loan or credit card. When we do some prescreening of a mailing list, for example, we now have two other sets of potential applicants whose performance we may need to infer—those who were excluded from the mailing and those who were mailed but did not respond. This is particularly important when we try to build response models; see (Bennett, Platts, and Crossley 1996) and (Crook, Hamilton, and Thomas 1992b).

In many countries, there are regulatory and self-regulatory rules concerning the use of information for prescreening. In the U.K., for example, the credit bureaus adhere to the principles of reciprocity regarding the sharing of credit information among lenders. Within this, there are rules about what levels of information a lender can use, and these vary, for example, depending on whether the person being considered is an existing customer. The rules also vary depending on whether the credit facility being considered is a product they already hold or is a new product. All of this varies also depending on the level of information the institution is supplying.

In the U.S., credit reference information is more readily available and reveals more about a potential borrower. On the other hand, restrictions are placed on a lender to the effect that, if they carry out a prescreen, an offer to a potential customer must then be an unconditional offer.

Another example of where we might use prescreening is if we have a response model and we can exclude from a mailing those who are unlikely to respond, thereby reducing the cost but not the effectiveness of the mailing. Indeed, it will generate a higher response rate.

Another factor to bear in mind with a response model or response scorecard is that it often works counter to a credit scorecard. At its simplest, those who are credit hungry are those who are most likely to respond, while those who are more creditworthy are less likely to respond. For example, we may find that younger people are more likely to respond to a loan offer or mailing, while older people are more likely to be deemed acceptable risks. The same might be true of other attributes. Therefore, we often have to reach a compromise so that we can attract sufficient volumes of applicants with a reasonable chance of being accepted and turning out to be good performers.

In a traditional scoring approach, our binary outcomes are that the customer will perform and not perform. With the first example of prescreening, this is also the case. We also discussed the binary outcomes of respond and not respond. Ultimately, when trying to maximize profit, the binary outcomes are composite ones of "will respond and will perform" and "will not respond or will respond but will not perform." (There are other occasions when we might use a prescreening approach. In one of these, we put applicants through an initial scorecard, and on the basis of the results, we could decline the application or approve it or carry out further assessment. Another is where the applications have to pass through a scorecard using only application data and a scorecard using credit reference data. In section 12.3, we discuss how we might combine two scorecards operated in such a fashion.)

10.3 Preapproval

Preapproval is prescreening taken to the next stage. Here we have carried out a sufficiently extensive prescreen that we are prepared to make the applicant a firm offer. This offer could be guaranteed or conditionally guaranteed, subject to some mild terms. One way to do this is with one's existing customers. If a loan customer has made their last 12 payments on a £5000 loan on time, one may be able to assess that they are a good risk and in control of their finances. One could even build a model that predicts the probability of such a customer later defaulting. In general, it is not too extreme to offer the customer another similar loan—say, for £5000 or £7500—but with a guarantee. Alternatively, one could make them a conditional offer of a much larger loan—say, £15,000—provided they could produce some evidence that they could afford it. Many lenders nowadays would not even require such evidence.

Another area in which this happens is with the management of credit limits on a credit card portfolio. In general, the higher the credit limit, the higher the balance. Therefore, credit card companies regularly review customer limits and amend them, usually upward, although if someone's performance has deteriorated, they may move the limit downward.

Many years ago in the U.K., the card companies reviewed the limits and then offered them to the cardholders and had to wait for them to reply stating that they wished the higher limit. Nowadays, the higher limits are effected without any customer interface. While discussing this, one might ask why a higher limit produces a higher balance. There are several reasons but one key factor to consider is that good cardholders do not want to be embarrassed when using their card. Therefore, they will use a card up to a margin below the limit. Once the balance reaches that point, they may start to use another card. If your card is the one they prefer, i.e., the one that is used first, the higher the limit, the longer it is before a second card is called into play. Thus in this way, a higher limit will not necessarily generate a higher expenditure from the cardholder, just concentrate that expenditure on one card.

10.4 Fraud prevention

In many ways, using scoring for fraud prevention is the same as any other use. We have experience of past cases and a binary outcome—genuine or fraud. The key difference is that good fraudsters will make their applications look very genuine. Therefore, some scoring developments for fraud prevention have not proved worthwhile because they are unable to differentiate between genuine applications and fraudulent applications. On the other hand, if we use scoring as a fraud check in addition to using a different scoring model as a credit risk check, any improvement, i.e., any success in identifying fraudulent applications, will add value. However, the value of this additional check relies on it not presenting too many false-positive cases.

Because of the specialized nature of fraud prevention, two other approaches may offer some help. These are neural networks and cross-matching. Neural networks were discussed in section 5.4, and an interesting application appears in Leonard (1993a, 1993b).

Cross-matching of applications takes a completely different approach and does not use statistical models. It works on the premise that once someone has been successful in perpetrating a fraud, they will attempt to repeat their success with another lender. Therefore, some lenders have begun to send details of applications into a central data bank, where some matching algorithms operate to identify common features. Many matching rules will be applied and it is acknowledged that many false-positive cases will be identified. However, this approach has also been useful in identifying common addresses and telephone numbers.

For example, an applicant uses 14 Main Street as their home address on a personal loan application but uses 14 Main Street as their lawyer's address on a mortgage application

and 14 Main Street as their employer's address on a credit card application. Even if these applications were made to three different lenders, only the first one would stand any chance of not being identified. Now it is possible, especially if the applicant is self-employed, that their home and work address and telephone numbers are the same. Also, if they are a lawyer, it is possible at they have named one of their associates as the lawyer for the transaction, and so the work and lawyer's address and telephone number are the same.

An applicant needs to use a genuine telephone number on an application in case the lender telephones as part of the application process. Also, in many cases, and even more so in the U.K. with checks to prevent money laundering, the telephone number will be checked against some directory. Therefore, if a fraudster has set up one or even a few dummy telephone numbers, these will be used repeatedly, either as home or business telephone numbers, and again cross-matching will identify these cases.

Scoring and other techniques for fraud prevention are continuing to be developed. The need for ongoing development arises because they have not yet reached a stage of offering significant chances of success. Also—and perhaps this is the reason for the previous comment—applicants attempting to commit fraud are changing their techniques and ways and, to some extent, are getting cleverer. However, given the annual amount lost on fraud, this is an area that will continue to attract attention.

10.5 Mortgage scoring

Scoring for mortgages has the same binary outcomes of performing and nonperforming. However, because it is for a mortgage, it may be quite different from scoring for a personal loan or a credit card in a number of ways, which we touch on in this section. First, there is some security or collateral—the property being purchased. Second, we may need to consider more carefully the interest rate schedule. Third, there is the eventual repayment to consider. Also, we need to consider different costs and processes—lawyers to take a legal charge over the property, valuations of the property, etc.

On a more positive note, scoring for mortgages may be simpler in some ways. Because there is a property secured that can be repossessed should the worst happen, we may be able to relax some of our concerns. We may be able to be less concerned about how the borrower is to repay the mortgage or whether the borrower will try to repay the mortgage, either on an ongoing basis or at the end of the term.

From as early as Chapter 1, we looked at what we are often trying to assess in application scoring. One aspect is the stability of the applicant, and one of the key factors or variables that might be used for this is the applicant's time at their present address. In mortgage borrowing, if a house purchase is involved—rather than a remortgage, where the borrowing moves from one lender to another while the borrower remains owning the same property—then we know that there is some instability involved as the borrower is about to move house. Therefore, if we wish to assess stability, we may need to do so using different means.

Another key factor in mortgages is the interest rate schedule and the income. Some mortgages are variable-rate mortgages. Here the rate varies and is set by the lender. The rate will be changed by the lender typically in response to a change in the interest rate set by the central bank. As a central bank lowers its rates, these reductions are, for the most part, passed on to the customer. Similarly, as rates rise, the mortgage rate will rise.

Some mortgages are fixed-rate mortgages. With these mortgages, the lender buys in funds from the capital market and lends it out to borrowers at a margin for a fixed term at a fixed rate. If the borrower continues with the mortgage until the end of its duration, the lender can assess the interest margin and income and the costs incurred at the application stage and

at the end of the mortgage. However, if the customer chooses to repay the mortgage at the end of the period of the fixed rate, or perhaps even earlier, the lender may not have enough income to cover the costs. To reduce the effect of this, many mortgage lenders impose restrictions on mortgage redemptions. For example, they may impose penalties if the mortgage is redeemed not only within the period of the fixed rate but also within a period after, known as the tie-in period. During the tie-in period, the mortgage interest rate applying to the mortgage will revert to some standard variable rate that typically will produce a wider margin for the lender. If the mortgage is redeemed during the period of the fixed rate, a penalty may apply, which is partly to recompense the lender for the funds that they have bought for the fixed-rate period and which they may now not be able to lend. Some approaches have been developed to address this, and in section 12.7 we look at survival analysis to model early repayment patterns.

There are other variations of mortgage interest rate schedules such as mortgages with capped rates, where the rate is variable but is guaranteed not to rise above a specified maximum. From a scoring point of view, the interest rate schedule affects our ability to decide if a mortgage is good or bad. This is partly because the profitability equation is altered dramatically. However, it is also affected by the fluidity of the market for mortgages and remortgages. In a fluid market, many mortgages do not reach a mature-enough point for us to decide whether they are good or bad.

While considering the profitability of a mortgage, it should also be realized that the asset and the repossession of the property in the event of default puts a notional, and perhaps actual, limit on the downside risk. Of course, the lender may also consider what to do at the end of the term of the mortgage. Specifically, should the lender insist on repayment or allow the customer to continue to make their regular monthly payments? The costs and the legal and valuation processes involved also affect the profitability of a mortgage. The proposition will need to consider who meets these costs and how effectively these processes have been carried out. Another factor that may have a bearing on the profitability of a mortgage portfolio is the use of mortgage-backed securities; these are discussed in section 14.6.

10.6 Small business scoring

If we can use scoring to assess lending propositions from consumers, then many people believe and almost as many have used scoring to assess lending propositions from small businesses. The key features are the same. We have a large volume of seemingly similar transactions. We can examine the information we had available to us at the point we had to reach a decision. We can review, with hindsight, which pieces of information would have allowed us to differentiate between loans performing well and those performing badly. Eisenbeis (1996) not only supported these views but also reviewed some of the modeling methodologies in use. However, there are many differences that should be recognized and considered.

Some of this information is subject to interpretation. For example, if a business produces audited financial accounts, these are produced for the specific purposes of taxation and statutory requirement at a given point and are out of date by the time an auditor approves them. In a very small business, it is up to the owners how much cash they withdraw from the business and how much the business retains. Also, whether the business makes an accounting profit or loss in a given year is not particularly indicative of the medium-term strength of the business and of its ability to generate cash and to repay any loan. The business may, in a normal course of events, change its strategy and its mix of operations. This can happen for even the smallest business. This may have an affect on our view of the likelihood of the business to succeed or to be able to support the lending being considered.

One part of the assessment of a lending proposition from a small business is an assessment of the owners. However, this may be a matter of interpretation. The person with the largest share in the business may not actually be the person running the business. (As we move up the corporate scale, the people who run large businesses are not the people who own them. These businesses are mostly owned by investment and pension funds.) With a small business, we need to define who actually affects the success of the business or is affected by the success of the business, because it is their stability and propensity to repay that we should try to assess.

In looking at a small business proposition, there is a greater need to consider the business environment. For example, a highly profitable and well-run business making circuit boards, say, may run into trouble if a major customer who is a PC manufacturer is hit by falling sales or adverse currency movements. However, once we consider these and many other factors, the conclusion is still that, with a lot of care and proper definition of subpopulations that are genuinely similar, it is possible to use scoring for small business lending.

10.7 Risk-based pricing

In risk-based pricing, we move the focus away from risk and closer to profitability. Risk-based pricing—or differential pricing, as it is sometimes called—is where we adjust the price or interest rate we offer to the customer to reflect our view of their risk or potential profit to us. While the binary outcomes continue to be whether the loan will be repaid, we now segment our population into different levels of expectation of performance or profitability and apply a different price to each. (Of course, it is possible that we adjust items other than the price. For example, we may reduce our requirement for security or collateral. In effect, this makes the proposition riskier but may be an acceptable business alternative to lowering the price for a low-risk proposition. We can also adjust requirements for documentary proofs, e.g., to substantiate income. With mortgage lending, we can also adjust our requirement for a property valuation to require either a more stringent or a less stringent valuation or none at all.)

In principle, we calculate the risk attributed to a proposition or customer and use this to assess the income necessary to make a profit or to achieve a target return. We then offer the customer the product or service at the interest rate to produce the required income. In effect, it means that very good customers will get to borrow money at lower rates and riskier customers will be charged a higher rate.

This is contrary to the standard way of doing retail business in western society. When we shop, the price of an item is displayed and either we choose to purchase at that price or we choose not to. From the lender's perspective, risk-based pricing usually increases the ability to sell loans and also means that the income more closely matches the expenses, at least in terms of the risk and the bad debt costs. However, there are also some potential challenges with this:

- Adverse selection. In scoring, analysis is carried out on past experience. Therefore, a consumer can be made different offers by different lenders. With risk pricing, this is even more widespread. The analysis to support risk pricing will include some assumptions about the take-up rate for marginal customers. However, we need to consider the fact that some or most of those marginal customers take out a loan at a higher rate because rates offered by others lenders are even higher. For example, if our standard loan is at an annual interest rate of, say, 12% and we offer marginal customers a loan at 14%, those who take up the offer are not a random sample of those

offered. Rather, they will be on average higher risk since the 14% offer will be lower than other offers they have received. To accommodate this, we should increase the rate offered, but this only aggravates the situation and we could end up chasing our own tail. Understanding the marketplace and the competitive environment is just as important when pricing to risk than with a standard offer.

- Good customers. At the low-risk end, we may be making offers at interest rates lower than the customer would deem acceptable. Therefore, for low-risk propositions, we may be undercutting ourselves and throwing away income. Again, an understanding of the marketplace and of the competitive environment is important.
- One needs to be able to explain risk-based pricing to customer-facing staff and to applicants. The main question that usually requires an answer is, "How did you calculate the rate you are now offering?" This is especially the case when the rate being offered is substantially different from an advertised rate.

Models for risk-based pricing are discussed later in section 14.5.

10.8 Credit extension and transaction authorization

We have referred to behavioral and performance scoring at several points. At this point we can explain some of the scenarios where we might use it. When an applicant applies for a credit card, credit or application scoring will be used to make the decision to issue a card, and the credit score may also have some impact on the actual credit limit to be granted. This credit limit is, however, an initial credit limit, and this limit needs to be managed throughout the lifetime of the account.

Once the account as been running for a few months and once some activity has been seen on the account, the customer's performance on the account gives us a much stronger indication of the ongoing likelihood of the customer's failure to operate the account within the terms set out. Even if we were to update the application data, the performance data will still be more powerful. So in what types of situation might we wish to use the score?

As stated earlier, the credit limit needs to be managed. If the limit is too low, the customer will concentrate their expenditure and usage on one or more other credit cards. Of course, if the limit is too high, then the customer may be tempted or encouraged to achieve a level of activity that they cannot afford to maintain. The skill of the area of credit limit management is to walk this tightrope between no income and high bad debts. Therefore, credit card companies will regularly review the credit limits of their cardholders and increase those that appear to be able to maintain a slightly higher level of expenditure.

There also will usually be an ongoing assessment that generates a shadow credit limit. This is not known to the customer but is in the background. It can be used if the customer either goes overlimit or is referred for a transaction that would take them overlimit.

Now, a credit card product is merely one example of a revolving credit type of product. In the U.K., there are other types of revolving credit. One is the overdraft, i.e., the account with a checkbook and a credit (overdraft) limit. In a similar way, such accounts will have a shadow limit so that when a customer requests an overdraft limit or a higher limit or when a check is presented that takes them over their current limit, the system can decide what action to take. Another revolving credit account is a budget account, where the customer receives a credit limit equivalent to 24 or 36 times an agreed monthly payment. Mail-order facilities will also involve some credit limit so that as a customer or an agent builds up their history, they are also able to build up the value of goods they can deal with on a credit basis.

We may also decrease credit limits. As the customer is not aware of their shadow limit, this can fall as well as rise. For example, should a credit card customer miss a payment, this will tend to reduce their score and therefore reduce their shadow limit. If adverse behavior continues, the credit card company may reduce the actual customer's credit limit and inform the customer accordingly. This can also happen with an overdraft. This may be effective in reducing the lender's exposure to the customer. It may also generate a change in customer behavior that, in many cases, is what is desired. Of course, in some cases it may be too late and is merely reducing or even only capping the eventual loss.

We may also use the behavioral score to allocate strategies for account handling. Once behavioral scores are available, we can move into the area of adaptive control. This involves managing accounts differently according to a number of key factors and learning from their behavior. Therefore, if a customer misses a payment, we may react differently to this depending on their score, the age of the account, the security held, and the past payment behavior. Many other factors can be used, but the principal aim is to manage accounts both effectively and efficiently.

For example, based on score, we may be able to predict that the customer who misses a payment after making the previous 12 payments is highly likely to catch up with their payments in the next month. Therefore, one possible course of action is to do nothing and wait for them to resolve the situation. Should the probability of their recovery be slightly less, another option is to send them a letter, and we can utilize a whole range of letters within this general strategy, from the soft to the hard reminder. An even lower probability may generate a collections telephone call. Of course, at the other extreme, should it be highly unlikely that the customer will recover, it may be that no action should be taken as it would be a waste of time and money. This last strategy, however, is rarely implemented.

With a revolving credit facility, the decision process may differ depending on how the credit application arose. For example, we may have different processes for credit limit increases requested by the customer and those generated automatically. We may have another process for credit limit increases generated by a customer writing a check or attempting a credit card transaction that takes them over their credit limit. Each different process could simply be a different allocation of score (and other variables) to limit. Alternatively, it could be a different scorecard if a large enough data set is available and supports such a difference.

10.9 Debt recovery: Collections scoring and litigation scoring

In the previous section, we touched on using the score in deciding what collections activity to carry out. This is part of behavioral scoring and, as mentioned above, the usual way to implement this is by means of adaptive control, where the lender implements various strategies and through them tries to learn what is most effective and efficient.

At a later stage in the process, a scoring approach has occasionally been used to decide what action to take. The available options include writing off the account, pursuing it, taking it through the legal process, sending it to a debt collection agency, and selling it to a third party. The outcomes in which we are interested are the likelihood of recovering the money. Of course, these are not simple binary outcomes. For example, one course of action may have an 80% chance of recovering all of the money and a 20% chance of losing a further 25%, say, to cover costs. Another course of action may have no chance of recovering all the money but is almost certain to recover 75% in each case. Both courses of action expect to lose 25% of the balance, and so there may be other issues involved in deciding which is the better course of action, such as resource availability or costs.

10.10 Provisioning for bad debt

Bad debt provisioning is not really a feature of lending. Rather, it is a feature of our accounting systems and of the control mechanisms laid down by the central banks and other regulatory bodies. It arises because it is deemed important that a lender has sufficient funds set aside to cover the bad debts that are likely to occur. However, the scale of these losses may be larger than expected, or the timing of them may cause them to occur sooner than expected. Funds set aside should be able to cover these as well.

When we ultimately finish with the debt collection and litigation process, we may have a balance remaining that we write off our books. A provision is, in essence, a prediction of a future write-off. Therefore, once the provision has been made or set aside, we do not need to concern ourselves with the timing of the eventual write-off since funds have already been set aside to meet this eventual loss. (This assumes that we have made a sufficiently large provision.)

In the U.K. environment, we refer to general provisions and specific provisions. At a simple level, specific provisions are those that are set aside for specific cases of bad debts that have already been recognized. For example, most lenders would set aside a specific provision once an account becomes three payments in arrears, and many will set aside provisions earlier than this.

A general provision is a provision set aside for cases of bad debts that we have not yet recognized but which we believe are in the current lending book. Suppose that we lend a new tranche of business with a score distribution such that we expect 2% of them to go bad. Suppose further that we expect 30% of the balances on the bad accounts to be recovered. We could set aside a general provision of 1.4% of the value of this new tranche of business as a general provision. Since we do not know yet exactly which accounts will fall into the 2%, we cannot set a specific provision. As these bad cases become evident, we can reduce the general provision and increase the specific provision accordingly.

How much we set as a specific provision and when we set this aside may be based on scores, especially behavioral scores, where we can distinguish between cases that are two down or three down but are more or less likely to recover.

10.11 Credit reference export guarantees

In the U.K., there is a branch of the government called the Exports Credit Guarantee Department (ECGD). It is not the function of this department to lend money. Its function is to support export trade, and one of the ways it does this is by guaranteeing an exporter that it will be paid. For them to understand their liability under such a guarantee, they need to assess the risk that the ECGD are running, and they do this with a form of scoring. Part of the scoring algorithm will be a factor related to the country into which the goods are being imported. In principle, however, the scoring methodology used is similar to those already discussed in this chapter. The ECGD's decision is in the form of a risk price, i.e., a cost for providing the guarantee.

This page intentionally left blank

Chapter 11

Applications of Scoring in Other Areas

11.1 Introduction

In the previous two chapters, we dealt with how scoring is used to make lending decisions. In this chapter, we look at some other areas where a scoring-type approach has been used but that are either at the periphery of lending or fall completely beyond the boundaries of a lending environment.

The first two topics have some relation to lending. We first deal with how we might use scoring in direct marketing. Then we look at profit scoring. After that, we move further away from the lending environment and touch on auditing in a variety of guises and then onto the parole process. The common features are that in almost all of these cases, there are two possible actions—analogous to whether to lend money—as well as data available on past cases—both at the point when a decision has to be made and ultimate performance.

11.2 Direct marketing

The previous chapter addressed direct marketing under the heading of prescreening. However, it is worth revisiting the topic to cement some of the ideas and to expand on the topic.

In many environments—and the lending environment is only one example—there are, in principle, two possible actions. Either we make someone an offer or we do not. (In practice, of course, there are many possible actions as we can choose from a range of offers or products, and a score could determine which offer to make.) A score might predict who is likely to respond and who is not. A score might also predict who is likely to be loyal and who is likely to move to a better offer from another organization as soon as one appears. A score might also predict who is likely to trade up to a better product. It should be obvious that these three examples would use different scores, i.e., different models.

We might also be able to build a model and produce a score to predict which marketing channel would be better to use. For example, we might be able to segment our target population into several segments according to whether we should market them by mail—cold mailing—or, for existing customers, statement inserts, or telephone, or even e-mail. (One useful model for such a development would be a multinomial logit model.) Clearly, the score becomes part of a larger business decision because it is much more feasible to mail 250,000 prospective customers than to make telephone calls to them, especially when one considers the necessity for repeat attempts to make contact with people unavailable on the first attempt.

If we are scoring prospects from our own list of customers or from a mailing or membership list, then there are clear restrictions. For example, we are limited to the information available. This may seem an obvious point, but if you haven't actually received a loan application yet, you cannot rely on purpose or term of loan as a scorecard characteristic (although you could make only a specific offer, e.g., for a 48-month loan for a car). Also, there are strict rules and regulations about what credit reference information is available and how it might be used. If the decision is whether to target someone for an offer, if it is decided not to select someone, they will not even be aware that they have been considered.

Thus far in this section, nothing is specific to lending. Indeed, models and a scoring methodology can work, in general, where we are trying to predict someone's propensity to do something. It may be that we are trying to offer them a visit to a timeshare complex and wish to predict, among a list of prospects, who is most likely to attend and to purchase. Similarly, we may be interested in offering people a chance to test drive a new model of car. Here we may be interested in those people who will make a purchase, whether they take up the opportunity offered. The clear message here is that there is no reason why a scoring methodology will not work well.

However, there are two warnings. First, with many uses of scoring in direct marketing, it is often the case that a scorecard trying to predict response will work in the opposite direction from one trying to predict sales or even performance. For example, the people who are most likely to respond to a loan offer are often those hungriest for credit and are least likely to repay satisfactorily. For example, the people most likely to respond to a cheap weekend at a timeshare resort as part of a sales promotion may be the people who are least likely to buy or least likely to be able to afford to make a purchase.

The second warning is that we should also consider the benefits of preselection. In a lending environment, if the target population is of fairly good quality, it may be cheaper to mail to all of them and to live with the few bad cases that apply and are approved than to prescreen all of them to remove the few bad cases. This is especially so once we realize that many of the bad cases would not apply anyway. In a sales environment, it might be better to make a blanket offer of a test drive rather than deselect some people, removing them from the prospect list. This might be the case where the cost of a test drive is minimal.

Whether we use scoring in lending or in direct marketing of other goods and services, the usual objective is to maximize some measure of profit. Often, the stated objective for the marketing department is response rate or cost per response. However, in terms of running a business overall, profit is the key measure.

11.3 Profit scoring

In most areas of scoring, we consider two outcomes—good and bad. Sometimes these are polarized by removing from consideration those cases in between—indeterminates. Often, good and bad refer to some performance definition. However, in most lending organizations, the ultimate objective is profit. Therefore, while we generally use performance definitions to classify accounts as good or bad, some scoring developments or implementations classify accounts as being good or bad depending on the profit they make.

In principle, this is simple. An account that is expected to make a profit is good and one that is expected to lose money is bad. In this section, we discuss two different aspects of this simple statement. The first is what to include in the calculation of profit or loss and the second is what profit measure to use. Both of these will affect scorecard development and implementation, although they rely on the introduction of some basic concepts in finance. (One or two of the finance concepts were discussed in section 8.11, where we discussed using profit as a means of deciding where to set the scorecard cutoff.)