Credit Scorecards for SME Finance
The Process of Improving Risk Measurement and Management

April 2009

By Dean Caire, CFA

Most of the literature on credit scoring discusses the various modelling techniques used to develop and validate scorecards. In contrast, this article focuses on the use and management of credit scorecards, regardless of how they were originally developed. We believe that any reasonably powerful scorecard adds value by providing a consistent measure of risk that can be used to improve business processes and inform other management decisions regarding loan approval, pricing, provisioning, and collections.

All Scorecards Are Not Created Equally, But They Are Treated Equally After Creation

Application credit scorecards are used to measure a prospective customer's credit risk—that is, the likelihood that the customer will repay his/her credit obligations. The measure, or score, allows us to rank clients by their risk. The ranking, or relative risk, in turn allows us to differentiate loan terms or service for clients by risk group. Finally, a consistent risk measure allows us to numerically estimate the impact of business decisions, such as tightening or loosening of credit policy, on future profits and/or human resource requirements.

There is a great deal of literature on the technical methods used to develop credit scorecards and measure their predictive accuracy. What all of these methods have in common is that they take information from the past—for example, historic portfolio data, market-pricing information, and/or the experience of senior credit officers—to identify and assign weights to indicators according to their association with higher or lower credit risk.

Regardless of how a scorecard is developed, we can evaluate its performance using an appropriate metric, identify its strengths and weaknesses, and select an appropriate strategy for its prudent usage. At this point, assuming we have developed the best scorecard possible given our data and resources constraints, all scorecards are again "equal" in that they must be monitored, periodically validated, and adjusted or re-developed as appropriate. All scorecards are also "equal" in the important sense that the scorecard equation itself is now just one of the pieces of a necessarily more complex credit business process.

In fact, once the scorecard has been developed, whether fully in-house or with the assistance of a third party, the long-term success of a credit scoring project will depend not only on how well the scorecard ranks risk or estimates a given applicant's probability of default (PD), but will also be a function of some or all of the following factors:

- Scoring's role in the business process, and the business process itself
- Software used to implement and administer the scorecard, including its links to other process-management software
- Training, support, and communication with "front-line" users
- "Ownership" of the scorecard by sufficiently senior people in the organization
- Regular monitoring of scorecard performance, along with readiness to adjust or re-develop the scorecard as appropriate
- Clear documentation of scorecard development and the scorecard validation process

Application Credit Scorecards and Basel 2

Credit scorecards traditionally estimate a probability that the borrower will not repay his loan. Since the Basel 2 accord was first published, the most common definition of credit risk has become the probability of default (PD) over the 12-month period following the application, or evaluation, date. Basel 2 has also spurred financial institutions to develop models of expected loss given default (LGD) and exposure at default (EAD). However, the Basel 2 accord does not specify any one methodology for the development of credit scorecards. Instead, it establishes some overriding principles for internal rating systems such as:

- Internal estimates of PD, LGD, and EAD must incorporate all relevant, material and available data, information and methods.
- Estimates must be grounded in historical experience and empirical evidence, and not based purely on subjective or judgmental considerations.
- The population of exposures represented in the data used for estimation, and lending standards in use when the data were generated, and other relevant characteristics should be closely matched to or at least comparable with those of the bank's exposures and standards.
- Internal ratings and default and loss estimates must play an essential role in the credit approval, risk management, internal capital allocations, and corporate governance functions of banks using the IRB approach.
  (International Convergence of Capital Measurement and Capital Standards, June 2006)

While the Basel accord calls for a degree of transparency and analytical rigour in model development and validation, it gives bank regulators in each country the job of 'approving' the use of internally developed rating models for provisioning.

In summary, the scorecard is not "complete" once it is tested out of sample, nor "perfected" when we exceed some benchmark statistic for model accuracy. Instead, the scorecard is more like one living cell in a larger, complex, credit-process-management organism. The scorecard and its use should grow and change over time in harmony with the larger organism. And, just as important, users and management should only use a scorecard while awake—that is, in combination with vigilant awareness of current economic and market conditions. As conditions change, the scorecard provides a consistent measure of credit risk to which we can adjust lending policy and find the correct balance between risk appetite and business targets.

In Scoring, Something Is Better Than Nothing, And Not Much Worse Than Something Better

Since all credit scorecards require ongoing monitoring and validation, the actual test statistic measuring its relative strength, such as the AUC (area under the curve), CAP (cumulative accuracy profile) or Gini coefficient, is practically relevant only during scorecard development, when it is one of the measures used to select the best of competing models. This includes choosing between a given model and no model, or what is called the "random model", and which really means lending without any formal scorecard. Our experience is that even in a worst-case modelling scenario, a scorecard developed with no data other than the knowledge of experienced credit analysts is far superior to the random model, i.e. no model, and therefore would give us a reasonable starting point for measuring and monitoring portfolio risk. Beyond this starting point, the goal will be to improve the risk model over time through regular monitoring and reporting, appropriate adjustments to the model or the policies for its use, and the collection of more and better data for future modelling/scorecard redevelopment.

If a scorecard is not the "silver bullet" that on its own can make accurate and depersonalized loan decisions, as it seldom is for SME loans, the question then is: how and where does a credit scorecard add value in the credit process? A scorecard that reasonably ranks applicants from low to high risk can bring the following improvements:

- Streamlined, quicker approval procedures for applicants in the lower risk categories.
- An increase in approval rates with stable or decreased delinquency rates.
- Risk-based segmentation of the portfolio for establishing credit policy: for example, grant new loans only to the lower risk clients in certain sectors.
- Risk-based pricing: charge higher prices to riskier borrowers.
- Risk-based/differentiated provisioning, assuming bank regulators approve the scoring model.
- Prioritization of regular monitoring and, particularly with delinquent loans, focusing available resources on riskier clients.
- Better data collection and storage as a result of the introduction of software to implement the scorecard. Often the introduction of scoring also coincides with a financial institution's first attempt to automate the credit process.

In other words, the introduction of a scoring system promises an immediate improvement in the management of data, and the ranking of risk according to one consistent measure opens up possibilities to increase operational efficiency in underwriting and collections. These benefits are present as long as the scorecard can reasonably rank risk and are not necessarily dependent on the sophistication of the method used to develop the scorecard

But How Can We Believe In What We Cannot Validate?

How a scorecard is validated depends on how much data is available for its development and testing. In the best-case scenario, there is enough data to validate the model out-of-sample, or, in other words, to fit the model to one set of data and test its predictive accuracy on another set of data. In this case, we can develop confidence intervals for the PD estimates and expect the model to perform within that degree of accuracy for as long as we believe current applicants and economic conditions resemble those in the period for which we had historic data. Greater confidence in model estimates allows us, for example, to prudently use the model more aggressively in recommending approvals and rejections and to estimate the impact of risk-based pricing on expected profits more precisely.

In the SME borrower segment, particularly in less developed or smaller credit markets (i.e., outside of North America and Western Europe), we often have to work with a scarcity of historical data, particularly with very few problem loans. In such cases, we cannot apply techniques such as logistic regression for model development, and particularly we cannot validate the scorecard out-of-sample. Instead, we conduct an "expert validation" by comparing the model rankings to the subjective assessment of experienced credit analysts, checking whether their assessments roughly "match", at least on a very crude scale of "low", "medium" and "high" risk. The mechanics of this exercise will depend on what types of rating or classification systems are already in place and the structure of the new scorecard, but the result should be that the analyst opinions relatively closely match the scorecard's rankings, particularly in the low and high risk tails. Given a model that approximates the credit assessment of experienced credit officers, many financial institutions will feel confident relying on the model more heavily for the best and worst cases and gradually reducing the in-between "gray" zone, of cases that require a fuller, standard review.

Moving Forward: Universal Reports For Scorecard Performance And Stability Monitoring

Once a scorecard is implemented, the ongoing monitoring and validation process will be nearly the same regardless of how the scorecard was originally developed. There are a few reports that are useful for monitoring the model's ability to rank risk and for evaluating in what ways the applicant population is changing over time. We present these report templates below[1].

The Delinquency-by-Score Report

This report shows the concentration of "bad", or non-performing loans, across possible score ranges. An example of the most simple delinquency-by-score report is shown below in Table 1. A model that accurately ranks risk should, over time, classify a progressively increasing share of delinquent loans (i.e. over 90 days in arrears) in score bands that indicate higher risk. In Table 1, a higher score indicates lower risk, and the model appears to rank risk fairly well since the bad rate gradually increases as the scores decrease.

Table 1: Delinquency-by-Score Report

A	B	C	D	E
			B+C	*C / D*
Score Range	Number "Goods"	Number "Bads" Over 90 Days Past Due	Total Number of Loans	Bad Rate
90-100	49	0	49	0.00%
80-89	167	0	167	0.00%
70-79	254	2	256	0.78%
60-69	488	5	493	1.01%
50-59	389	5	394	1.27%
40-49	291	7	298	2.35%
30-39	166	8	174	4.60%
20-29	88	6	94	6.38%
10-19	42	3	45	6.67%
0-10	0	0	0	0.00%
TOTAL	1,934	36	1,970	1.83%

In practice, when portfolio volumes are small or during the first year of model implementation when there is limited default experience, the "Bad Rate" may not always consistently increase as scores decrease. Over time, however, the pattern of problem loans should indicate that loans with higher scores have lower bad rates.

[1] Examples of similar reporting templates with more detailed explanations can be found in Grigoris Karakoulas (2004) Empirical Validation of Retail Credit-Scoring Models. *The RMA Journal* and Naeem Siddiqi (2005) Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring. *Wiley*.

Slightly more complex calculations performed on the information presented in Table 1 are used to decide on the best place for cut-off points and to measure the model's cumulative power (such as the AUC, CAP, or Gini statistics mentioned above). For choosing cut-off points, rather than using a particular formula , we suggest looking at the number of good and delinquent loans in each score band and then calculating, as precisely as possible, the expected economic gain or loss of approving or rejecting all the applicants above or below a given cut-off.

In Table 2, we present the data from Table 1 with the additional column (G) indicating the percentage of loans with a score in a given scoring band (column A) or higher. For example, 2.49% of loans score 90 points or higher, 10.96% of loans score 80 points or higher, etc. If we were to automatically approve all loans scoring above 60 points in Table 2, we would approve nearly 50% of applicants (965/1,970 =0.4898) and expect a bad rate of less than 1%, or (2+5)/(965) = 0.00725. If we can put a numeric figure to the cost savings per loan of an automatic or scorecard-based approval and also estimate the costs of each bad loan we will accept, we can estimate the economic gain or loss of approving all loans scoring 60 points or above.

Table 2: Delinquency-by-Score Report with Additional Information on Cumulative Distribution

A	B	C	D	E	F	G
			B+C		*C / D*	*E / 1,970*
Score Range	Number "Goods"	Number "Bads" Over 90 Days Past Due	Total Number of Loans	Cumulative Number of Loans	Bad Rate	% of Loans with Score Greater Than or Within Score Range
90-100	49	0	49	49	0.00%	2.49%
80-89	167	0	167	216	0.00%	10.96%
70-79	254	2	256	472	0.78%	23.96%
60-69	488	5	493	965	1.01%	48.98%
50-59	389	5	394	1,359	1.27%	68.98%
40-49	291	7	298	1,657	2.35%	84.11%
30-39	166	8	174	1,831	4.60%	92.94%
20-29	88	6	94	1,925	6.38%	97.72%
10-19	42	3	45	1,970	6.67%	100.00%
0-10	0	0	0	1,970	0.00%	100.00%
TOTAL	1,934	36	1,970		1.83%	100.00%

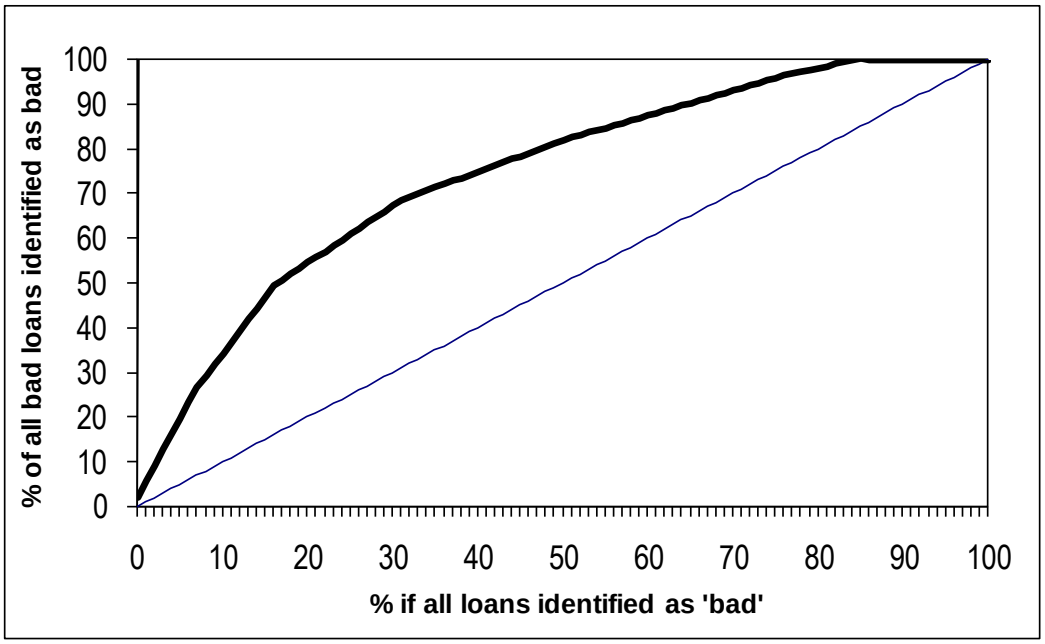
As an example of selecting a rejection policy, we could estimate the effects of rejecting all loans below a certain cut-off score. Looking again at Table 2, even if we automatically reject only the lowest scoring loans, those scoring 19 points or less, in the past we would have rejected 45 clients, of which 93% (42/45 = 0.9333) were in fact good customers. Unless the cost of working

out a bad loan is very high, it is very unlikely that it makes economic sense to use the model in this example for automatic rejections at all. Instead, in such cases where models are weak in the high-risk tails, we would recommend subjecting all loans below some cut-off point (60 points in this example) to additional subjective review.

Figure 1 illustrates the "area under the curve" we are analyzing. Without focusing on the AUC statistic itself[2], we can visually interpret the model's strengths and weaknesses by examining the shape of the heavy black line, our model, as compared to the random model, or no model, shown by the thinner diagonal line. The closer the heavy line is to the left vertical axis, the better the model can isolate bad loans in high risk score zones, as in the example model in Figure 2. The closer the heavy line is to the top horizontal axis, the better the model concentrates 'good' loans in the low risk score zone, as shown in the example model in Figure 3.

Our example in Figure 1 does a little better on the low risk side, as we have already discussed, but this is, in our opinion, easier to observe by analyzing Table 2, which we also find most informative for setting scorecard decision policy. The numerical AUC statistic would be one of several factors that could be used to select between competing models.

Figure 1: Area Under the Curve for Example Model



(Chart Template: Microfinance Risk Management (2009) www.microfinance.com)

Figure 2: Area Under the Curve, Model that Better Isolates 'Bads' in High Risk Zones

[2] A fuller explanation of the Receiver Operating Characteristic, including discussion of the AUC statistic, can be found at: http://en.wikipedia.org/wiki/Receiver_operating_characteristic

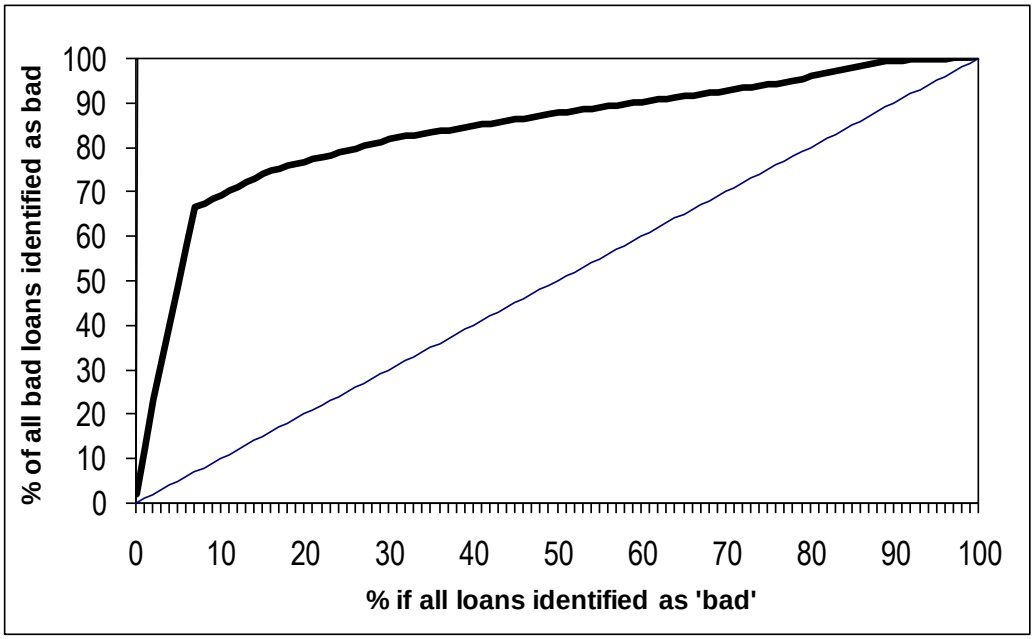
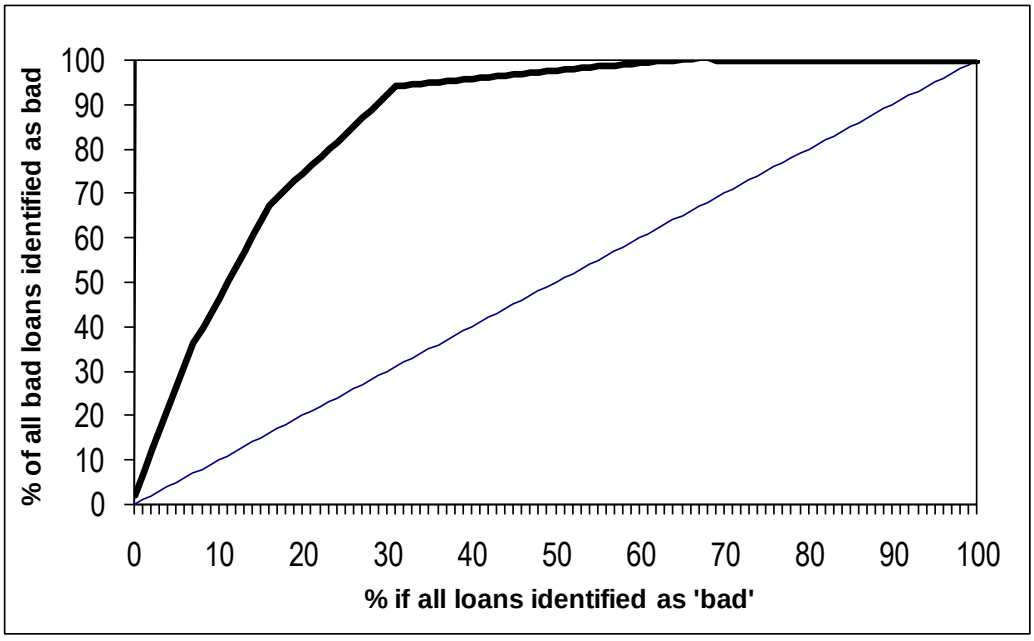


Figure 3: Area Under the Curve, Model that Better Isolates 'Goods' in Low Risk Zones



Interpreting Back Testing Data Gathered in a Different Lending Process

When a report like Table 2 is used to analyze back-testing data, or data on loans that were not originally scored but evaluated using a different process, we should be more careful extrapolating the historic delinquency-by-score to future expected performance. Our experience with scorecards for the SME segment is that we can expect a delinquency-by-score report generated with back-testing, or historic, data, to be applicable to a new scoring process if the following elements are common to both the original underwriting process and the new scoring process:

- Fraud checks, such as verification of applicant's identity and registration documents, checks of credit registries and 'black-lists'
- A checklist of minimum lending criteria, sometimes called "stop factors" – when any of these minimum criteria is not met, the loan must be reviewed more carefully or elevated to a higher approval competency
- a visit (or not) to the client premises

When the above steps were included in both a standard and a scoring loan approval process, the scorecard itself comes to replace a more in-depth analytical process with a single numerical, objective measure. Once we validate that the scores "agree" with past experience and/or credit analyst judgment, we can be confident that, with some margin of error, we can estimate future performance based on the scorecard's performance on past data.

The Population Stability Report

Another important aspect of monitoring credit scorecards is to check whether recent applicants are similar in terms of risk characteristics to the applicants whose experience we modelled. The most effective way to do this is to compare the distribution of scores of more recent clients to the distribution of scores from our model development and testing data. A Population Stability Report, as shown in Table 3, presents the changes in the distribution of scores between the most recent period and the model development data. Because the distribution of loans by score range is stated in percentage rather than absolute numbers, such a report can be used even when the model was initially developed only with expert judgment and tested on a small sample of applications, provided the data for this sample was recorded.

Table 3: Population Stability Report

A	B	C	D	E
			B-C	*B/C*
Score	Actual %	Expected %	Actual – Expected	Actual / Expected
90 – 100	2.49%	0.00%	2.49%	0.00
80 – 89	8.48%	11.02%	-2.54%	0.77
70 – 79	12.99%	13.00%	-0.01%	1.00
60 – 69	25.03%	25.92%	-0.89%	0.97
50 – 59	20.00%	15.51%	4.49%	1.29
40 – 49	15.13%	12.81%	2.32%	1.18
30 – 39	8.83%	11.81%	-2.98%	0.75
20 – 29	4.77%	5.70%	-0.93%	0.84
10 – 19	2.28%	4.23%	-1.95%	0.54
0 – 10	0.00%	0.00%	0.00%	0.00

Column B in Table 3 presents the observations in each score band as a percentage of all observations in the most recent period.
Column C shows the same information for the observations in the development data set. Columns D and E measure the changes in the score distribution between the most recent period data and model development data.

Figure 4 presents the same population stability information (columns B and C of Table 3) graphically. Analysis of either Table 3 or Figure 4 would lead to the conclusion that current applicants appear to have similar risk profiles to the model development applicant pool.

Figure 4: Population Stability Graph, No Significant Change

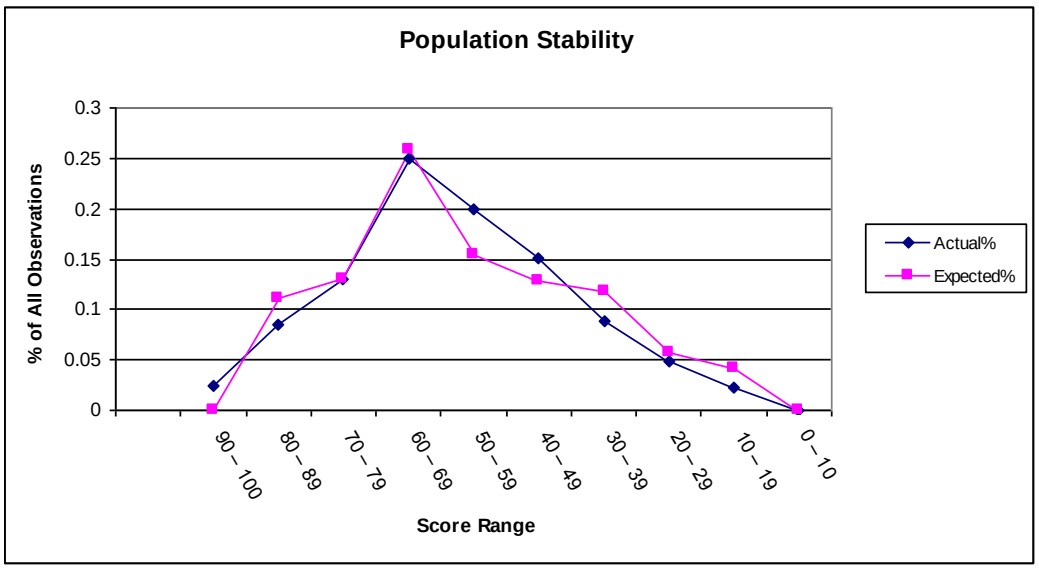
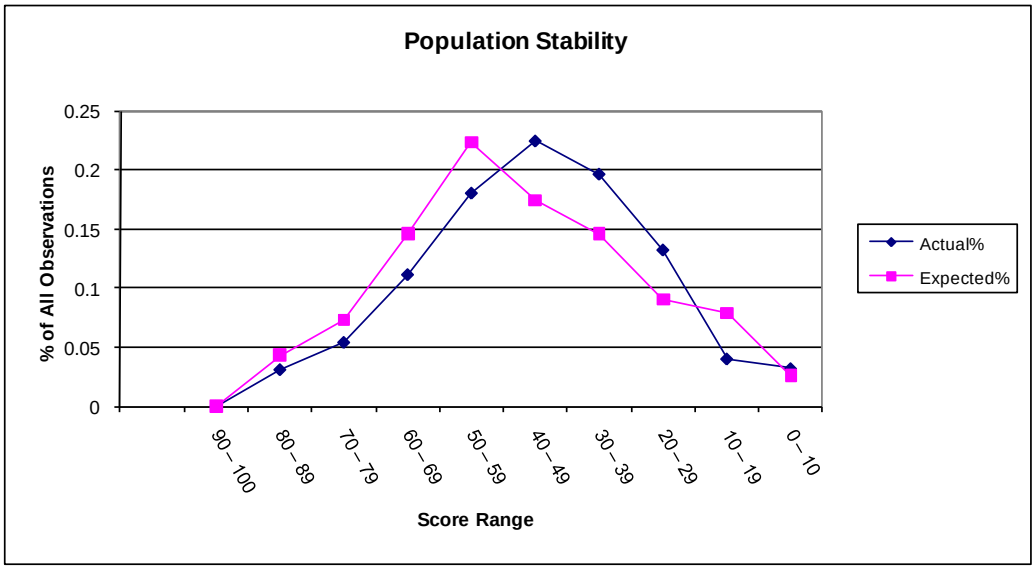


Figure 5, drawn from the same "Expected" data set, but with a different set of "Actual" data, , by contrast shows a pronounced right-ward shift in the "Actual %" score distribution, indicating that current applicants, as measured by our scorecard, are on average more risky than applicants in the development data sample. Although it is possible to calculate a change index, we do not recommend focusing on any particular index number, but instead interpreting the overall shape of the change and then looking deeper into changes on a factor by factor basis using Characteristic Analysis reports to answer the question "why are average scores lower for our recent applicants?"

Figure 5: Population Stability Graph, Increase in Risk



Characteristic Analysis Reports

Characteristic Analysis reports help to pinpoint what is causing any changes in total score distributions over time. As shown in Table 4 below[3], this report calculates an index (in column E) as the difference between the expected (column B) and actual (column C) score distributions per response category multiplied by the points per response category (column D). The total index figure is a sum of the individual response categories and is interpreted as the number of points, higher or lower, that recent clients are scoring, on average, for that characteristic.

[3] The template for this report was taken from Naeem Siddiqi (2005) Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring. *Wiley*

Table 4: Characteristic Analysis for Credit History Variable

A	B	C	D	E
				*(C-B)*D*
Credit History	**Expected**	**Actual**	**Points**	**Index**
No previous loans	46.85%	66.67%	0	0.00
Over 30 days past due	2.95%	3.11%	-30	-0.05
15 - 30 days past due (Not > 30 days)	9.65%	8.20%	5	-0.07
8-14 days past due	1.97%	2.11%	20	0.03
Always current-7 days past due	38.58%	19.91%	40	-7.47
	100.00%	100.00%		**-7.56**

Table 4 shows an example of a Characteristic Analysis report for the variable Credit History. We can see that clients in the most recent ("Actual") period are scoring, on average, 7.5 points less than the clients in our model development sample. By looking closely at columns B and C, we can see that a larger percentage of recent clients have no previous loans (66% percent as compared to 46% in the development sample). This also means that a lower percentage of clients have regular payment discipline of "Regular 0-7 days" (19% vs. 38% in the development sample), and the index (column E) shows us that the 7.5 point difference can be traced back to this response category. The most general conclusion could be summarized as: "today's clients are less likely to have a credit history."

Generally, characteristic analysis will help to identify in what ways the applicant population has changed, but there are no rules of thumb for how to respond to changes. If changes are very large, such that it appears our recent applicants no longer resemble the applicants whose past repayment behaviour we modelled, we should redevelop the scorecard. When changes are more subtle, it could make sense to adjust the weights for certain response categories.

To conclude this section, we would like to stress that the types of simple management information reports presented above should be generated and analyzed for all scorecards, regardless of whether they were developed using logistic regression or only expert judgment. While the appropriate method for scorecard development depends on the quality and quantity of data available for model building, the practical usage of scorecards, including ongoing monitoring and management, is the same for all scorecards.

When To Redevelop A Scorecard?

There is not one answer to the question: when is the best time to redevelop a scorecard? Some common "triggers" for re-development or adjustment are:

1. **There is a significant improvement in the quality or quantity of data**. For example, an expert model is initially developed due to lack of data. After 18 months, assuming there have been some delinquencies of over 90 days (or whatever other definition is selected to discriminate between good and problem loans), logistic regression can be used to test the current model factors, select other potential factors, and optimize scorecard weights.

2. **The Delinquency-by-Score report indicates that the model is not ranking risk correctly**. Regardless of the original development approach, re-development or significant adjustment could be appropriate. However, we again will mention that these adjustments should be made only after there have been some non-trivial number of delinquencies, so that the report will show patterns, rather than the result of a few particular cases, some of which could have gone delinquent due to factors besides credit risk -most frequently this would be fraud.
3. **The Population Stability and/or Character Analysis reports indicate that recent applicants are significantly different from the applicants whose experience was used to specify the scorecard**. This re-development trigger is also irrespective of original development methodology and potentially could be addressed through adjustment rather than full redevelopment.

Conclusion

In summary, scorecard development is a very important first step in the long-term success of a credit scoring project, but it is only that—a first step. The success of a scorecard project depends on active management, regularly monitoring and periodic adjustments, based on which the quality of the scorecard's risk measurement and its role in risk management can only improve and expand.

Dean Caire is a Risk Management Specialist for DAI Europe, a global consultancy working with financial institutions worldwide to advance human prosperity. In case of questions, please contact: dean_caire@dai.com.

Special thanks to Mark Schreiner of Microfinance Risk Management and Mary Miller of DAI for editing this paper.