Empirical Validation of Retail Credit-Scoring Models

by Grigoris Karakoulas

Despite its emphasis on credit-scoring/rating model validation, Basel II does not impose any standards on the process. This article presents a methodology for validating credit-scoring and PD models, focusing on retail and small business portfolios. A future article will address corporate portfolios.

Always a good idea, development of a systematic, enterprise-wide method to continuously validate credit-scoring/rating models nonetheless received a major shove from Basel II. As we've all come to know, the Accord requires qualifying banks to have robust systems for validating the accuracy and consistency of rating systems and processes. Further, banks need to be able to estimate the risk components, namely, probability of default (PD), loss given default (LGD), and exposure at default (EAD).

The validation process also is important for corporate governance purposes. It can help detect deterioration in a model's performance, which could affect existing risk-tolerance limits and economic capital allocation. The process can also assist in maintaining the loss/revenue objective associated with the implementation of a scoring model.

This article presents a methodology that can serve both purposes—validating credit-scoring models used for customer adjudication and validating the estimation of the risk components. Application and behavior scores may be used as input for pooling retail portfolios as well as for estimating the risk components.

According to a 2003 ISDA-RMA survey, the range of available data has caused banks to employ a range of validation techniques, resulting in key differences in the techniques used for corporate versus retail portfolios. This articles focuses on credit-scoring models for retail and small business (typically less than $200,000 credit) portfolios.

Methodology

Credit-scoring models are usually static in that they do not account for the time to delinquency or default and are built from a couple of point-in-time snapshots. There are various reasons that could cause actual performance of a scoring model to deviate from its expected performance—that is, performance at the time the scoring model was developed. For example, a scoring model might lose its predictive power during a recession if the characteristics entered into the model or the underlying customer population are sensitive to the economic cycle. In such cases, the distribution of the input characteristics could shift. Also, a scoring model may continue to rank-order the population and provide acceptable discriminant power, yet fail to produce desired performance be-

 Grigoris Karakoulas is an adjunct professor in Computer Science at the University of Toronto and president of InfoAgora, Inc.

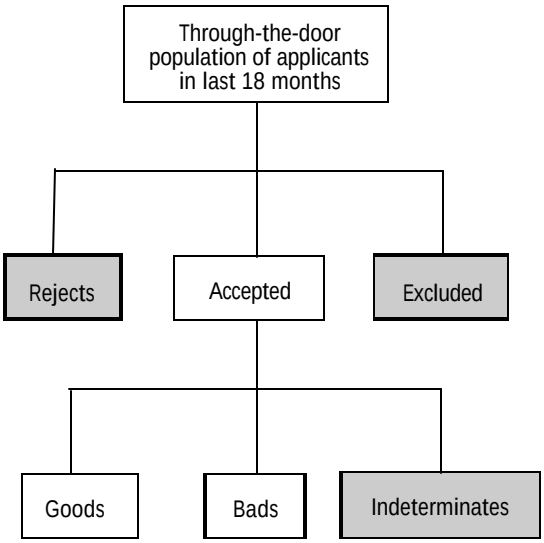
cause the scores (probabilities) from the model have lost their calibration with respect to the actual probabilities in the current population. If cutoff scores are used for adjudication, adjustments of those cutoff scores may be necessary.

Three diagnostic techniques for monitoring the performance of credit-scoring models can help us check for deterioration. The first technique can detect shifts in the score distributions of the development and current populations. The second technique can detect changes in the rank-ordering power of the model. Both techniques are presented with tests for assessing statistically significant changes. The third technique can be used to explain any possible misbehavior identified from the application of the first two techniques, by examining the characteristics input to the model as potential causes of that misbehavior. The application of the third technique is therefore conditional on the outcome from the first two techniques. In combination, these three techniques enable us to build an early warning system for detecting deterioration in credit-scoring models.

The first step is to define the data required for the validation process. Figure 1 shows an example of such data. For a given credit product, collect data for the through-the-door population of applicants over the past K months (for example, K=18). The data includes the model characteristics, adjudication outcome, and, as required, the credit performance of approved applicants. Accurate comparisons between the current and development populations require us to use the same data definitions as those used at the time of model development.

Based on these definitions, filter out any applicants that were excluded at development, such as applicants who were manually adjudicated. Then divide the population into accepted/rejected, and divide the accepted applicants into "goods," "bads," and "indeterminates." The latter group could be, say, accepted applicants with fewer than six months' performance history or, in the case of revolving products, accepted applicants whose credit remains uncashed.

Figure 1
A Breakdown of the Through-the-Door Population of Applicants in Last 18 Months ("current" population)



Think of the through-the-door population and its good/bad subpopulation as "current," to be compared with their counterpart populations at the time of model development ("development" populations). At this point, there are three steps in the validation process.

Step 1: Shifts in Score Distribution

Figure 2 shows an example of model misbehavior. The score distribution in the through-the-door population of a credit product should be stable over time. To detect any shifts in that score distribution, calculate the population stability index as follows:

- Bucket the population into score bands that are equally distant: for example, 50 score bands for a score in the 0-100 range, each band being 2 points wide.
- For each score band, calculate the number of applicants and the number of accepted and rejected applicants in the current and development through-the-door populations. Table 1 shows a tabulation of such numbers.
- Given the applicant counts, you can calculate the probability of each score band in the current and development populations.
- Measure the difference between the resulting two distributions using the *population stability index*. This is a measure of the distance between two probability distributions. A value of the index above 0.25 indicates a significant shift in the score distribution. It requires characteristic analysis (see Step 3) for understanding why this shift occurred.

Figure 2

Shift in the Score Distribution of the Development and Current Through-the-Door Populations

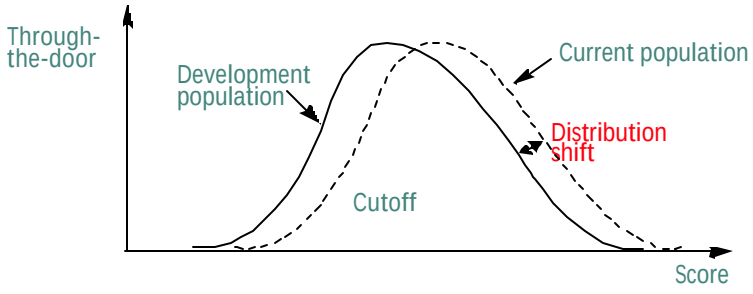


Table 1

Population Stability Index from a Sample Scoring Model; the Lower Scores Correspond to Higher Risk

			Current						Development			
Score Band	Population Count	%	Accepts Count	%	Rejects Count	%	Population Count	%	Accepts Count	%	Rejects Count	%
0-30	46	0.2	4	0	42	0.9	167	0.5	42	0.2	125	0.9
31-32	133	0.5	7	0	126	2.8	544	1.6	99	0.5	445	3.3
33-34	91	0.3	3	0	88	1.9	303	0.9	76	0.4	227	1.7
35-36	129	0.5	2	0	127	2.8	246	0.7	85	0.4	161	1.2
37-38	101	0.4	37	0.2	64	1.4	142	0.4	73	0.4	69	0.5
39-40	99	0.4	2	0	97	2.1	100	0.3	100	0.5	0	0
41-42	237	0.9	18	0.1	219	4.8	778	2.4	119	0.6	659	4.9
43-44	308	1.1	4	0	304	6.7	915	2.8	226	1.2	689	5.1
45-46	746	2.7	30	0.1	716	15.7	1224	3.7	421	2.2	803	5.9
47-48	1264	4.6	676	2.9	588	12.9	1337	4.1	634	3.3	703	5.2
49-50	221	0.8	164	0.7	57	1.3	178	0.5	109	0.6	69	0.5
51-52	310	1.1	21	0.1	289	6.3	1036	3.1	379	2	657	4.8
53-54	129	0.5	5	0	124	2.7	448	1.4	144	0.7	304	2.2
55-56	635	2.3	436	1.9	199	4.4	1476	4.5	611	3.1	865	6.4
57-58	4528	16.4	4097	17.8	431	9.5	4616	14	2394	12.3	2222	16.4
59-60	4375	15.8	4080	17.7	295	6.5	2624	8	1649	8.5	975	7.2
61-62	89	0.3	4	0	85	1.9	100	0.3	45	0.2	55	0.4
63-64	60	0.2	3	0	7	0.2	90	0.3	3	0	6	0
65-66	35	0.1	21	0.1	14	0.3	24	0.1	4	0	10	0.1
67-68	27	0.1	23	0.1	4	0.1	43	0.1	23	0.1	20	0.1
69-70	31	0.1	23	0.1	8	0.2	120	0.4	55	0.3	65	0.5
71-72	103	0.4	86	0.4	17	0.4	250	0.8	147	0.8	103	0.8
73-74	798	2.9	753	3.3	45	1	1463	4.4	923	4.8	540	4
75-76	4390	15.9	4165	18.1	225	4.9	8696	26.4	6037	31.1	2659	19.6
77-80	5182	18.7	4978	21.6	204	4.5	1235	3.7	1230	6.3	105	0.8
81-82	200	0.7	100	0.4	100	2.2	300	0.9	300	1.5	0	0
83-84	100	0.4	100	0.4	0	0	100	0.3	100	0.5	0	0
85-86	400	1.4	400	1.7	0	0	600	1.8	500	2.6	100	0.7
87-100	2909	10.5	2826	12.3	83	1.8	3827	11.6	2898	14.9	929	6.8

Validation Dates	Stability Index
June 2004	0.21(X)
March 2004	0.1 (X)

Green level (X): 0-0.10
Orange level (X): 0.10-0.25
Red level (X): greater than 0.25

While the above technique has been applied to a scoring model used for adjudication, it can also be applied to a model that estimates the PD risk component for Basel II, since it can identify any PD calibration issues for that model by comparing the PD distributions of the development and current populations.

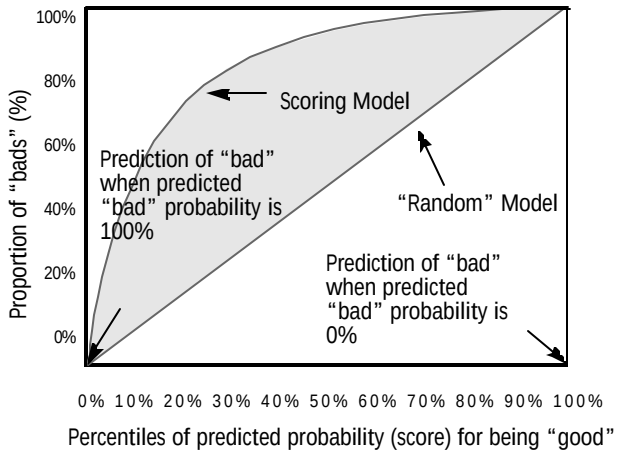
Step 2: Changes in Ability of the Model to Rank-Order

A model should continue to rank-order the accepted applicants (accounts) and also discriminate between goods and bads as it did in the development population. Let us assume a model that produces a 0-100 score—the higher the score, the better the account. If you rank-order the account from the bad to good score, a perfect model would rank all the bads at score 0. This is an optimal situation that corresponds to perfect separation of the goods and bads distribution. In practice, models score most of the bads close to 0.

The degree of how much a model scores a bad account closer to 0 than a good account can be measured by the Mann-Whitney U statistic that counts 1) the number of times a bad account has a score less than a good account's and 2) half the number of times a bad account has a score equal to a good account's. This statistic is graphically depicted by the area under the curve in Figure 3, called *area under the ROC* curve (AUROC). The diagonal line in the figure corresponds to a model that randomly assigns scores, hence called the *random* model. A *scoring* model should be doing better than the random model, that is, the gray area in Figure 3. When comparing two areas under

Figure 3

ROC Graph for a Scoring Model



the ROC curve, a Chi-squared test can show whether the difference between two values of the statistic is statistically significant at a particular confidence level, such as 95%.

It is worth pointing out that the commonly used K-S statistic is the maximum distance between the ROC curve and the x-axis. However, this maximum distance may occur at any point in the ROC curve. It is better if it occurs at the low scores, where most of the bads should be, than at the high scores. The K-S statistic has the limitation that it does not refer to where the maximum distance occurs. The Mann-Whitney statistic for AUROC is more general and so is better than the K-S statistic. Related to ROC is the Lorenz (also called CAP or lift) curve and the associated Gini measure that is used by some banks for validation.

Use the current and development goods/bads populations to estimate AUROC. Then, apply the significance test on the difference between the current and development statistic. Depending on the result of the test, you may accept or reject the hypothesis that the difference is zero.

In addition to rank-ordering, a scoring model should maintain the calibration of its score to the odds of "bad." For a model with scores in 0-100, for example, the odds at the 60 score are 25:1. It is important also to look at the odds-to-score relationship because it is usually used for setting the cutoff score and thus affects the accept/reject rate. A scoring model may continue to rank-order, but its cutoff score might have to be adjusted if there is a calibration weakness in the odds-to-score relationship. Figure 4 shows an example of this relationship in the current and development good/bad populations. In this example, the scoring model maintains its positive slope. However, it has lost some of its calibration because the scores correspond to higher odds in the current population than in the development population. If this relationship change persists over two consecutive validation runs, then the cutoff might have to be adjusted to reflect the new relationship.

As with the population stability index in Step 1, the AUROC statistic can also be used for validating a model that estimates the PD risk component for Basel II.

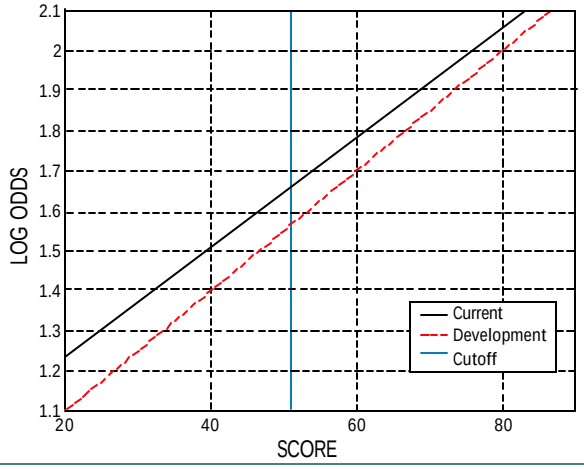
Step 3: Shifts in Characteristics Distributions

In cases where Steps 1 or 2 identify model misbehavior, a characteristic analysis should be applied to the current and development populations to identify which characteristics (attributes) could be the causes of that misbehavior. For each characteristic and value, the proportion of applicants should be generally the same in the current and development through-the-door populations. Any deviation in these proportions will affect the score output from the model.

The current and development distributions of a characteristic can be compared and the statistical significance of any shift can be assessed through:

Figure 4

Example of Score-To-Odds Relationship in the Development and Current Populations



- The Kolmogorov-Smirnov test if the characteristic takes real values.
- The Chi-squared test if the characteristic takes nominal values.

The magnitude of the effect from a shift in a characteristic distribution depends on the significance of the characteristic and its specific value in the calculation of the final score. If the scoring model is linear, as is typically the case with scorecards, then you can assess the effect of a shift in the distribution of a characteristic by multiplying the amount of the shift per value with the corresponding weight for that characteristic-value combination. Table 2 presents an example of a characteristic analysis summary for a linear scoring model. Here, the shifts in the characteristics distribution result in positive contributions to the final score and, therefore, higher values of the score in the current population. If the scoring model is not linear, it may not be easy to assess the effect on the score from a shift in the distribution of a characteristic. To approximate the effect on the score, you could estimate the derivative of the score with respect to the characteristic and then multiply that estimate by the shift in the distribution of the characteristic.

Summary

The diagnostic techniques shown in this article allow a bank to build a system for early detection and diagnosis of any deterioration in the performance of its credit-scoring and PD models across all retail credit portfolios. The Basel II Accord requires producing validation studies on a periodic basis. This validation process should be run on each credit portfolio quarterly or semi-annually, depending on the number of new bad accounts that are incrementally available in that portfolio.

Of course, the question is when a scoring model should be adjusted or redeveloped if it fails in a validation run. The decision to adjust the cutoff score is easier to execute than the decision to redevelop the model. You may wish to defer model redevelopment until failure over two validation runs, depending on the severity of the failure and the frequency of the validation runs. Severe deterioration and/or a model that starts performing more like a random model may require immediate redevelopment.

If a scoring model is used for adjudication, you might also want to perform a cost-benefit analysis for the redevelopment of the model. To do this, you can use the cost of a false positive (opportunity cost from a false "reject" obtained through reject inferencing) and a false negative (actual loss from a false "accept") and compare the total cost from the performance of the current scoring model with the cost of developing and implementing a new scoring model.

There are different techniques to validate models that estimate the LGD and EAD risk components, since the underlying problem is a regression one. Those techniques will be addressed in a future article. ❐

Contact Grigoris Karakoulas by e-mail at grigoris@infoagora.com.

References

1. International Convergence of Capital Measurement and Capital Standards: A Revised Framework, Basel Committee on Banking Supervision, Bank for International Settlements, June 2004.

2. Internal Ratings Validation Study: Report of Study Findings, ISDA-RMA, 2003.

Table 2

Example of Characteristics Analysis Summary from a Linear Scoring Model

Characteristic	Overall Score Difference	
	June 2004	March 2004
Net worth	-1.32	-0.57
Number of trades high ratio	0.45	0.05
Number of inquiries	2.30	2.00
Number of trades 60 days delinquent	1.50	1.25
Number of trades open in last 12 months	0.33	0.20
Percent of trades delinquent	1.30	1.12
Months in file	1.07	0.87
Months since most recent delinquency	0.57	0.24
Net fraction revolving burden	-0.44	-0.34