Validating a Credit Score Model in Conjunction with Additional Underwriting Criteria

September 2012

INTRODUCTION

Model validation is a critical activity to verify that credit scorecards are working as intended and that model usage is in line with business objectives and expectations. A regular model tracking and validation process can ensure that consistent and optimal model-based decisions are being made. It can also serve as an early warning system for identifying when a change may be necessary, whether it be an adjustment to a score cutoff strategy or a full model redevelopment. The process can be a straightforward exercise when one model is being used in isolation.

However, prudent lenders do not rely exclusively on a score alone to make credit decisions. Often data that is not available to the credit scoring model can provide additional predictive power, and a well-designed overlay strategy can optimize the decision-making process by utilizing all of the available predictive data. While an integrated strategy can provide superior results to a one-model solution, common model validation procedures may no longer be appropriate for the ongoing tracking and measurement of the credit scoring model. This paper provides a methodology for validating the credit scoring model when it is being used in conjunction with overlay criteria.

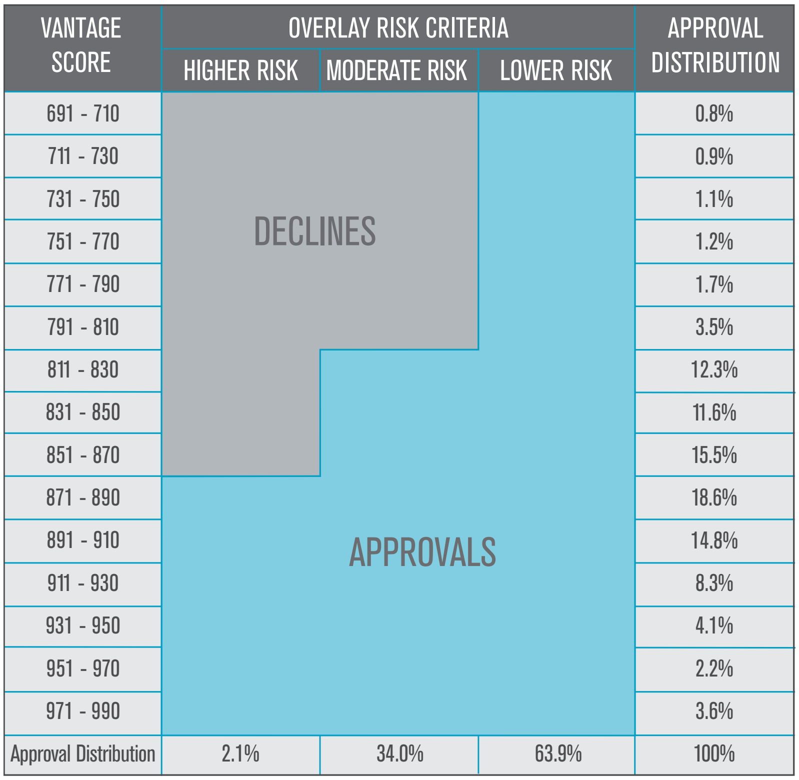
AN EXAMPLE

There are many reasons that a lender might augment a credit score with additional data for decision-making. In underwriting new loans, there is usually additional information contained in a credit application (e.g., income, debt-to-income ratio, employment status) that is not included in the credit score obtained from a credit reporting company. Some lenders are able to incorporate data from other customer relationships, such as product usage or length of relationship. Another common practice is to employ business rules that are specific to the product being underwritten, such as putting limits on the loan-to-value for certain risk levels. In all of these cases, the credit score is not the sole determinant of risk, and therefore the subsequent revalidation of the scoring model may provide counter-intuitive or misleading results.

Consider the following underwriting example in which a lender is using the VantageScore® credit scoring model in conjunction with a separate risk dimension. This “overlay” dimension can be thought of as an internal risk indicator or a set of alternative criteria that further segments the population into three levels of risk: “Higher,” “Moderate,” and “Lower.” The strategy for approving or declining applications is illustrated in Figure 1.

AN EXAMPLE
(Cont.)

FIGURE 1
SAMPLE UNDERWRITING STRATEGY UTILIZING OVERLAY RISK SEGMENTS



VANTAGE SCORE	OVERLAY RISK CRITERIA			APPROVAL DISTRIBUTION
	HIGHER RISK	MODERATE RISK	LOWER RISK	
691 - 710				0.8%
711 - 730				0.9%
731 - 750				1.1%
751 - 770	DECLINES			1.2%
771 - 790				1.7%
791 - 810				3.5%
811 - 830				12.3%
831 - 850				11.6%
851 - 870				15.5%
871 - 890				18.6%
891 - 910		APPROVALS		14.8%
911 - 930				8.3%
931 - 950				4.1%
951 - 970				2.2%
971 - 990				3.6%
Approval Distribution	2.1%	34.0%	63.9%	100%

This overlay strategy enables the lender to accept a limited number of loans with lower credit scores[1], provided they fall into the "Lower Risk" overlay segment. Similarly, a high credit score enables the lender to approve a small number of "Higher Risk" loans.

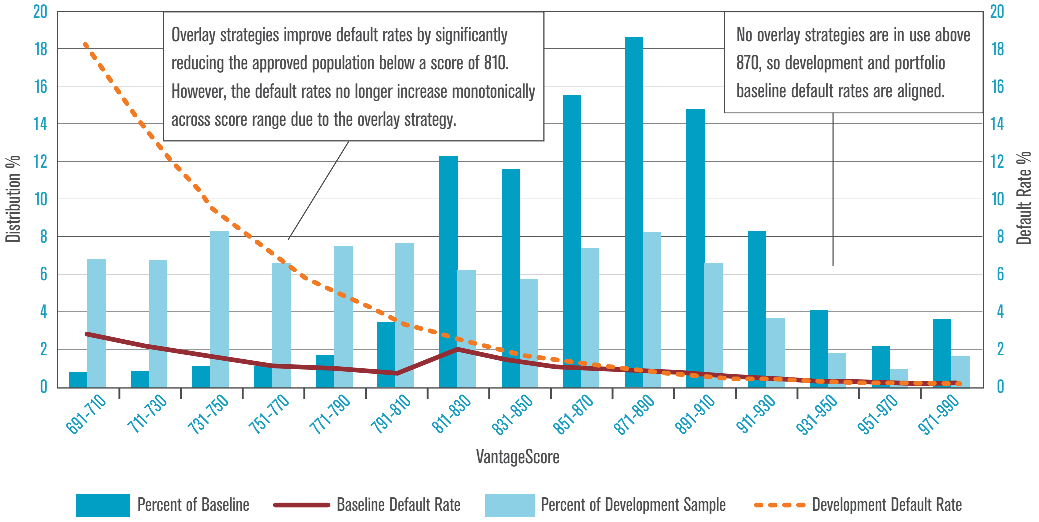
The first step in validating a credit score model is to establish the baseline default rates to which actual performance will be compared. The baseline provides a frame of reference for the validation results, and will be unique to a lender's portfolio and credit strategy. To obtain the baseline metrics for the sample underwriting strategy described above, the overlay matrix is applied to a set of loans for which performance is already known; this will generate the "expected" default rates that will ultimately be used in validating the credit score model.

[1] The VantageScore credit scoring model's range is 501 – 990.

AN EXAMPLE
(Cont.)

Figure 2 illustrates the impact that the overlay strategy has on overall model performance and how the baseline default rates can differ dramatically from model development estimates.

FIGURE 2
BASELINE VS. DEVELOPMENT DEFAULT RATES WITHIN VANTAGESCORE



Overlay strategies improve default rates by significantly reducing the approved population below a score of 810. However, the default rates no longer increase monotonically across score range due to the overlay strategy.

No overlay strategies are in use above 870, so development and portfolio baseline default rates are aligned.

In Figure 2, the dashed curve represents the default rates that could be anticipated if no overlay strategy were in place, and is representative of default rate estimates from a generic model development. The solid line indicates the baseline results and shows the impact of employing the overlay strategy; namely, that the overlay allows the lender to identify lower risk applications in the 691 to 810 score range that could not be identified by credit score alone. As a result, the default rate for these loans is lower than if identified solely using the credit score. For example, Figure 2 shows that the default rate for loans scoring between 711 and 730 is approximately 12%; however, after the overlay strategy is applied the default rate for approved loans drops to just over 2%.

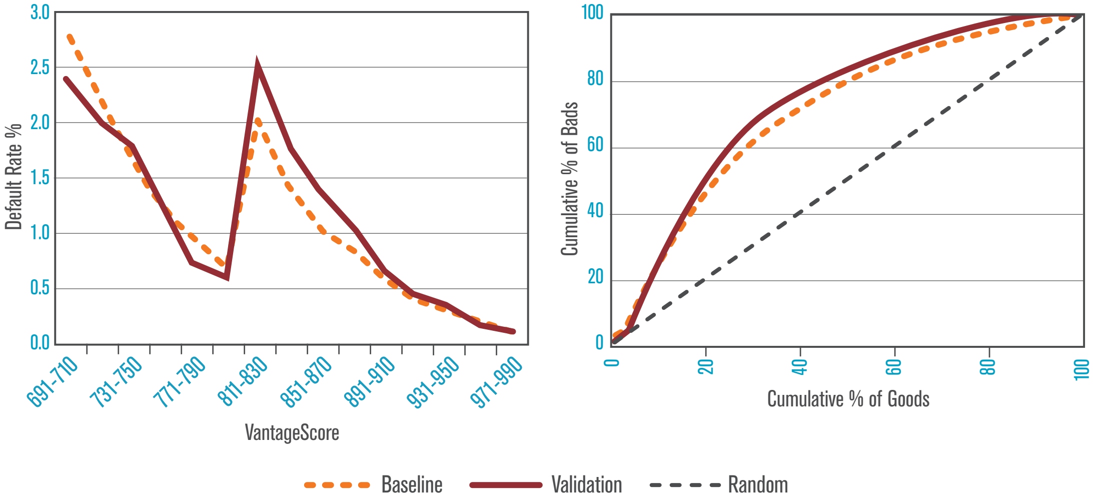
The baseline results also demonstrate that the resulting rank-ordering of default rates across credit scores is no longer a monotonic function, making it more of a challenge to determine whether the model is continuing to rank order risk effectively. Note that in Figure 2, the shape of the solid line is purely the result of incorporating the overlay strategy, and is not indicative of any scorecard performance issues.

VALIDATION CHALLENGES

When validating a credit scoring model, two vital areas of performance are *rank-ordering* and *separation*. Rank-ordering is assessed by computing “bad rates” (typically default rates) for different levels of the credit score model. Model separation can be examined graphically by plotting the ROC curve (or trade-off curve) as well as by computing statistics (such as KS) and comparing them to a prior time period or to an alternate model[2]. When an overlay strategy is present, the traditional approaches to model validation may no longer be sufficient in determining whether the model is performing appropriately. To illustrate this, Figure 3 shows aggregated model validation results for our example overlay strategy.

FIGURE 3

DEFAULT RATES BY VANTAGESCORE **ROC CURVE–BASELINE VS. VALIDATION**



The first chart shows default rates by credit score; the dashed line shows the baseline results and the solid line reflects the actual default rates during the performance window. Higher default rates are evident for scores between 811 and 890, but the presence of the overlay strategy makes it impossible to truly assess the rank-ordering of the model. The second chart shows the ROC curve for the validation (solid line) compared to the baseline (dashed line). Superior models will have curves that are pulled to the top-left of the graph. In this case, the chart suggests that we are seeing better model separation in the validation than in the baseline.

[2] A detailed discussion of model validation methodologies can be found in the *Executing Effective Validations* webinar at *www.vantagescore.com.*

PIECE-WISE VALIDATION

While the Figure 3 charts provide directional indications on model performance, more can be understood empirically by deploying the following method.

An effective methodology to validate a model in the presence of an overlay strategy is to perform a piece-wise validation on segments that have received consistent treatment. This will allow for a clean read of model performance, and will provide additional detail on how the model is interacting with the overlay strategy.

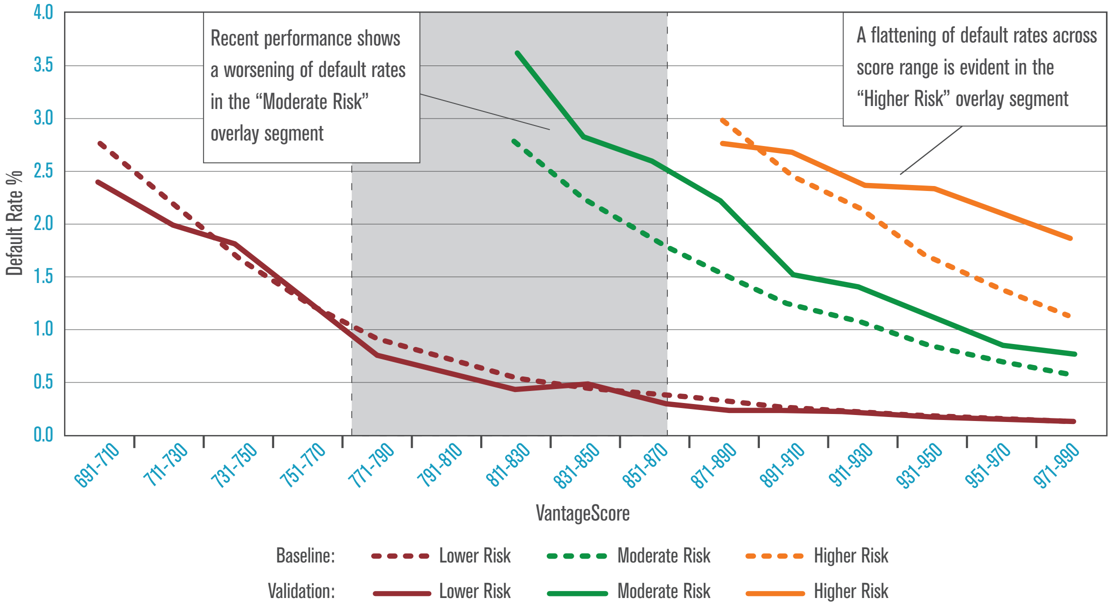
Using our earlier example, the lender has implemented different credit score cut offs for each of the overlay segments, as shown in Figure 4.

FIGURE 4

SEGMENT (OVERLAY)	VANTAGESCORE RANGE ACCEPTED
Lower Risk	691-990
Moderate Risk	811-990
Higher Risk	871-990

Model performance will be truncated for the "Higher" and "Moderate" risk segments because loans in these segments are subject to more stringent credit score requirements. Since the lender's underwriting strategies are consistent *within* each of the overlay segments, it is appropriate to evaluate the score's predictive performance on a segment-by-segment basis. The default rate graph has been recreated below (Figure 5), now showing the actual results compared to the baseline for each segment separately.

FIGURE 5
DEFAULT RATES BY OVERLAY RISK LEVEL



Recent performance shows a worsening of default rates in the "Moderate Risk" overlay segment

A flattening of default rates across score range is evident in the "Higher Risk" overlay segment

PIECE-WISE VALIDATION
(Cont.)

When comparing the validation results to the baseline in Figure 5, there are three general areas on which to focus:

1. *Monotonicity* – A strong model will continue to effectively rank order default rates across the entire score distribution.

2. *Consistent range of default rates* – If the validation default rates cover a narrower range than the baseline, the model may no longer be providing enough separation.

3. *Stability of point estimates* – A model that maintains similar default rates over time is desirable, but significant changes from baseline default rates are not always the result of a poor model.

The separation of performance into the three overlay segments now provides clearer insight in terms of how the credit score model is performing. Monotonicity does not appear to be a problem, as the score continues to rank order default rates effectively in all three segments. In addition, the performance of the model on the "Lower Risk" segment is very strong, with default rates that are aligned closely to the baseline. However, the graph suggests that there are potential areas for review in the other two overlay segments.

The "Moderate Risk" segment is exhibiting substantially higher default rates than were observed in the baseline sample. The baseline default rates ranged from a low of 0.5% to a high of 2.8%, but the validation results are tracking higher and range from 0.7% to 3.6%. If this trend were evident in all three overlay risk segments, it might be indicative of an external factor like an economic downturn. In this case, the higher default rates are not portfolio-wide, so the segment itself must be analyzed. Although the rank-ordering of the model remains strong, the following research is suggested to determine the root cause of the higher default rates:

- The segment definition should be checked to see if it has remained constant from the time the strategy was implemented; a slight change in criteria could be responsible for the higher default rates.

- An audit should be performed to confirm that loans are being assigned to the correct overlay risk segment.

- The stability of the underlying segment population should be analyzed to identify any distributional shifts that could be responsible for the change in performance. Examples of these shifts may include: a spike in loans from higher risk geographies, a reduction in average borrower income or a change in the mix of loan durations. Any shift that has occurred could impact a segment's default rates, so the population mix of the baseline and validation populations should be compared to identify potential sources of performance differences.

Ultimately, the lender will likely need to take some action on this segment by either re-evaluating the "Moderate Risk" segment definition or adjusting the minimum score required for applications in this group.

The “Higher Risk” segment highlights a different problem: the narrowing of the range of default rates. This can be seen graphically in Figure 5 as a “flattening” of the default rate curve. A portfolio-wide flattening trend would be an indicator of model deterioration, and if significant enough, could point to the need for model redevelopment. However, since it is limited in this case to a segment comprising only 2.1% of loan approvals, the result may be due to the small sample size. The lender should still perform the same analyses recommended for the “Moderate Risk” segment to ensure that the strategy is working as intended.

In addition to the default rate graph, it is also instructive to look at the relative differences in the KS statistic between the baseline and the validation results to assess model separation. The KS values calculated at the segment level will be much lower than would be seen in a full population development. This occurs for two reasons. First, each of the segments has been truncated as a result of the minimum credit score requirement; a majority of the “bads” have been declined due to low credit scores, and thus have been excluded from the population available for analysis. The other reason is that as the population being scored becomes more homogeneous, any credit scoring model’s performance will exhibit diminished KS values. Although it is not appropriate to compare KS values *across* segments, it is valid to compare them *within* segments as shown in Figure 6.

FIGURE 6
RELATIVE KS DIFFERENCES, BASELINE VS. VALIDATION

OVERLAY SEGMENT	BASELINE KS	VALIDATION KS	% DIFFERENCE
Lower Risk	29.5	31.1	5.5%
Moderate Risk	13.5	12.0	-10.9%
Higher Risk	8.0	3.6	-54.6%

This KS table shows that the “Lower Risk” segment is actually experiencing a modest improvement in separation, while the “Moderate Risk” group is slightly worse. The change in KS for the “Higher Risk” segment confirms what was visually evident in the default rate graph.

ALTERNATIVE VIEW

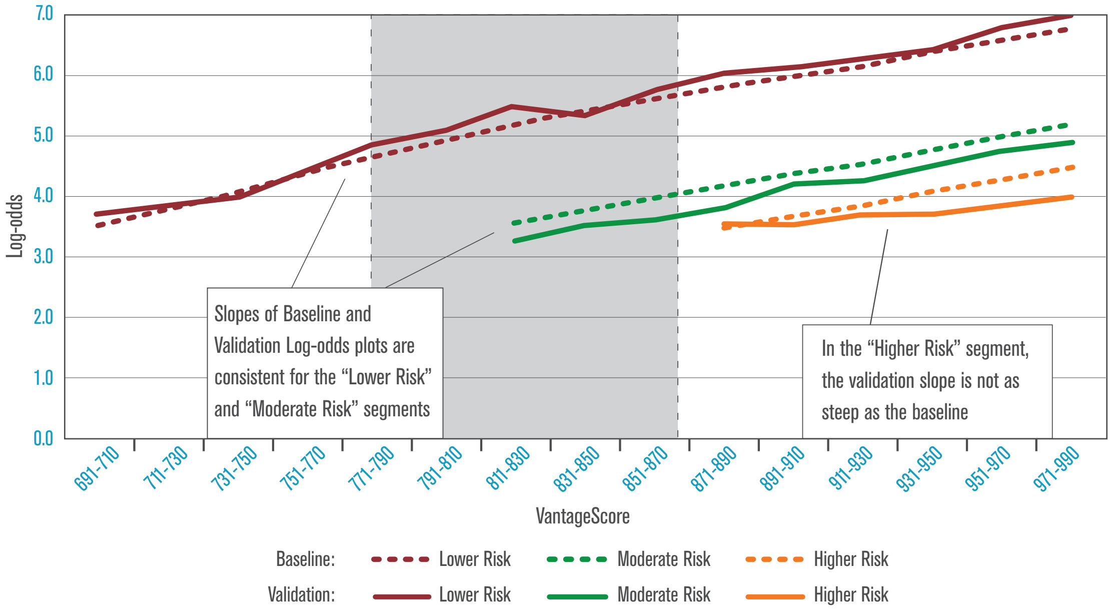
Another way to assess model performance is to graph the natural logarithm of the good-to-bad odds ratio (i.e., Log-odds) across score range for each overlay risk segment. Note that the same analysis is applied. The good-to-bad odds ratio can be calculated as follows: `(1 – Default Rate)/(Default Rate)`.[3] The logarithm function is then applied because most credit score models are developed having a linear relationship with Log-odds. This relationship makes it easier to identify when score deterioration may be occurring.

[3] In this formula all non-defaults are considered “goods.” If an “indeterminate” definition is also being used, the percent of indeterminates can be subtracted from the numerator for a more precise calculation.

PIECE-WISE VALIDATION
(Cont.)

Figure 7 interprets Log-odds in relation to overlay risk level. While the interpretation of this graph is similar to that of the default rate graph, the linear relationship between credit score and Log-odds enables us to calculate the slope of each of the plotted "lines."

FIGURE 7
LOG-ODDS BY OVERLAY RISK LEVEL



Slopes of Baseline and Validation Log-odds plots are consistent for the "Lower Risk" and "Moderate Risk" segments

In the "Higher Risk" segment, the validation slope is not as steep as the baseline

As with the KS chart, the validation results for the "Higher Risk" segment have changed dramatically from the baseline as seen in Figure 8. The larger overlay segments are exhibiting results that are consistent with the original baseline data, and do not indicate any model deterioration.

FIGURE 8
SLOPE OF LOG-ODDS, BASELINE VS. VALIDATION

OVERLAY SEGMENT	BASELINE SLOPE	VALIDATION SLOPE	% DIFFERENCE
Lower Risk	0.23	0.24	3.3%
Moderate Risk	0.20	0.21	2.5%
Higher Risk	0.20	0.08	-59.9%

VALIDATING THE OVERLAY SEGMENTATION STRATEGY

An additional step in the validation process when an overlay strategy is being utilized is **to validate the segmentation strategy itself.** One approach is to review the relative default rate differences between segments, controlling for credit score. The baseline analysis described earlier provided the data to compute an initial set of default rate "multipliers" that will be used to compare against the validation results. The multipliers are calculated by simply dividing the default rate for one overlay segment by the default rate for another segment *within the credit score band*. Figure 9 shows the default rate multipliers for our sample strategy and includes a column for each overlay segment combination. A multiplier for a pair of segments that remains within 10% of its baseline indicates a very stable segment relationship. If a multiplier differs from the baseline by more than 20%, the segment definitions should be examined carefully to determine the cause of the shift. Note that for our sample strategy, the overlay segmentation for credit scores below 811 cannot be analyzed directly, as only the "Lower Risk" segment had loans that were approved in these score bands.

FIGURE 9
DEFAULT RATE MULTIPLIERS, BASELINE VS. VALIDATION

DEFAULT RATE MULTIPLIER	MODERATE RISK VS. LOWER RISK	HIGHER RISK VS. LOWER RISK	HIGHER RISK VS. MODERATE RISK
Baseline	**5.0**	**10.0**	**2.0**
Validation:			
811 - 830	8.6		
831 - 850	6.0		
851 - 870	8.2		
871 - 890	9.3	11.9	1.3
891 - 910	6.9	12.3	1.8
911 - 930	7.5	12.9	1.7
931 - 950	7.0	14.6	2.1
951 - 970	7.8	19.4	2.5
971 - 990	8.0	20.1	2.5
Average	**7.7**	**15.2**	**2.0**
vs. Baseline	*+54%*	*+52%*	*-1%*

From this chart, we can see that the baseline default rates for the "Moderate Risk" segment were five times higher than those of the "Lower Risk" segment across all score bands. The validation shows that this segment now has default rates that average 7.7 times higher than the "Lower Risk" segment—an increase of 54% over the baseline. The "Higher Risk" segment exhibits a similar trend when compared to the "Lower Risk" segment—on average the default rate multipliers are 52% higher than the baseline. The last column of the chart shows that the relative difference between the "Higher Risk" and "Moderate Risk" default rates is consistent with the expectations from the baseline. As we have already observed that the default rates for the "Lower Risk" segment are in-line with the baseline results, this provides further evidence that the "Higher Risk" and "Moderate Risk" segments should be re-examined to determine if they are being implemented incorrectly or if the underlying population has shifted. Adjustments to the overlay strategy may be necessary for the lender to maintain acceptable risk and profitability levels.

CONCLUSION

When performing a model validation in the presence of overlay criteria, it is important to keep in mind that any metrics computed at the aggregate portfolio level will not be indicative of the true performance of the model. As outlined in this paper, an effective validation should include:

1. Establishment of an appropriate baseline—This will ensure that the overlay strategy is taken into account *a priori* and that aggregate metrics are put into the proper context.

2. Piece-wise validation of overlay segments—Treat each overlay segment as a unique portfolio when evaluating score performance, but remember that inconsistent performance across segments could point to a problem with the overlay strategy itself.

3. Overlay strategy analysis—Drilling down into the overlay segment definitions and underlying population stability will identify potential strategy inconsistencies.

Remember, while traditional methodologies and portfolio metrics may provide directional insight into model performance, the overlay strategy itself is an additional variable that must be accounted for in each step of the validation analysis.