Paper SAS2603-2015

Addressing AML Regulatory Pressures by Creating Customer Risk Rating Models with Ordinal Logistic Regression

Edwin Rivera, Jim West, and Carl Suplee, SAS Institute Inc.

ABSTRACT

With increasing regulatory emphasis on using more scientific statistical processes and procedures in the Bank Secrecy Act/Anti-Money Laundering (BSA/AML) compliance space, financial institutions are being pressured to replace their heuristic, rule-based customer risk rating models with well-established, academically supported, statistically based models. As part of their customer-enhanced due diligence, firms are expected to both rate and monitor every customer for the overall risk that the customer poses. Firms with ineffective customer risk rating models can face regulatory enforcement actions such as matters requiring attention (MRAs); the Office of the Comptroller of the Currency (OCC) can issue consent orders for federally chartered banks; and the Federal Deposit Insurance Corporation (FDIC) can take similar actions against state-chartered banks.

Although there is a reasonable amount of information available that discusses the use of statistically based models and adherence to the OCC bulletin *Supervisory Guidance on Model Risk Management* (OCC 2011-12), there is only limited material about the specific statistical techniques that financial institutions can use to rate customer risk. This paper discusses some of these techniques; compares heuristic, rule-based models and statistically based models; and suggests ordinal logistic regression as an effective statistical modeling technique for assessing customer BSA/AML compliance risk. In discussing the ordinal logistic regression model, the paper addresses data quality and the selection of customer risk attributes, as well as the importance of following the OCC's key concepts for developing and managing an effective model risk management framework. Many statistical models can be used to assign customer risk, but logistic regression, and in this case ordinal logistic regression, is a fairly common and robust statistical method of assigning customers to ordered classifications (such as Low, Medium, High-Low, High-Medium, and High-High risk). Using ordinal logistic regression, a financial institution can create a customer risk rating model that is effective in assigning risk, justifiable to regulators, and relatively easy to update, validate, and maintain.

INTRODUCTION

Assessing customer risk is an essential component of a comprehensive Bank Secrecy Act/Anti-Money Laundering (BSA/AML) monitoring program, because a financial firm's ability to identify customers that pose a higher risk for money laundering and terrorist financing is key to implementing effective customer due diligence (CDD) policies, procedures, and processes. As the Federal Financial Institutions Examination Council's (FFIEC) *BSA/AML Examination Manual* clearly explains, the concept of CDD begins with verifying each customer's identity and assessing both the specific risk and the risk level associated with that customer, as well as putting processes in place to ensure the additional scrutiny that higher-risk customers require.

As part of CDD, firms are expected to both rate and monitor their customers for the overall risk that they pose. Because of the increased regulatory emphasis in the BSA/AML community on using statistical processes and procedures to ensure compliance, banks and other financial institutions face pressure from regulators to stop using heuristic, rule-based customer risk rating processes and instead switch to academically supported, statistically based models. As these firms consider their model options, they must also take into account the Office of the Comptroller of the Currency's *Supervisory Guidance on Model Risk Managemen*t (OCC 2011-12), which outlines the modeling development process, validation requirements, and governance framework that must accompany the model.

Customer risk rating models are systems, algorithms, or processes that firms use either to assign customers to various risk groups or to score the customer's relative risk based on that customer's inherent characteristics, expected transactional behaviors, or overall status at the firm. These models usually fall into either of two primary types: heuristic, rule-based models and statistically based models. As firms look to improve their current customer risk rating models or to implement models where they currently don't exist, questions often arise, such as:

- What are the advantages and disadvantages of using one type of model versus the other?
- Why the regulatory push toward using statistically based models?
- What type of statistically based models should the firm implement?
- What attributes should the firm consider when developing the model?

HEURISTIC, RULE-BASED MODELS VERSUS STATISTICALLY BASED MODELS

Traditionally, financial firms have created customer risk rating models by focusing on rating customers' risk in several distinct areas, often with multiple variables in a single area. Here are some of the variable types that a customer risk rating model can include:

- *Customer Relationship* (personal, business, commercial, etc.)
- *Geography* (country of residence, business location, High Intensity Financial Crime Areas (HIFCA), High Intensity Drug Trafficking Areas (HIDTA), port or border cities, etc.)
- *Account Features* (remote deposit capture (RDC), correspondent banking, online banking, custodial accounts, etc.)
- *High-Risk Customer* (non-resident alien (NRA), politically exposed person (PEP), money service business (MSB), employee, etc.)
- *Alert / Filing History* (manual alerts created, system-generated alerts, Cash Transaction Reports (CTRs), Suspicious Activity Reports (SARs), etc.)
- *Expected Product Usage* (wires—domestic or foreign, cash, Automated Clearing House (ACH), check, etc.)
- *Expected Transactional Activity* (aggregate dollar amount of activity expected)

Although both heuristic, rule-based models and statistically based models consider the same basic set of customer data attributes, the underlying methodology that each type of model uses to weight the variables and score the customers differs—often significantly.

HEURISTIC, RULE-BASED MODELS

A heuristic, rule-based model is simply an analytical formula used to assign a score based on one or more variables, or attributes, that the firm deems important. Often these models are created using all available variables, because the relative importance of each individual variable is generally unknown and thus variable selection is quite difficult. As Figure 1 shows, these models are often parameterized so that you can adjust the scores and weights assigned to each component of the model.

Variable Type	Attributes	Logic Description
Customer Relationship	Customer Type	+ If the customer is “Personal,” then the score is 35. + If the customer is “Commercial,” then the score is 20.
High Risk Customer	Money Services Business (MSB)	+ If the customer is “MSB,” then the score is 80.
High Risk Customer	Politically Exposed Person (PEP)	+ If the customer is “PEP,” then the score is 80.
Alert / Filing History	Suspicious Activity Reports (SARs)	+ If the SAR count equals 1, then the score is 45. + If the SAR count is greater than 1, then the score is 60.
Expected Transaction Activity	Total Aggregated Transactions	+ If the monthly transaction volume is less than $50K, then the score is 40. + If the monthly transaction volume is greater than or equal to $50K, then the score is 60.

Figure 1. Example of a Heuristic, Rule-Based Customer Risk Rating Model

Each individual customer’s scores are then aggregated, and the firm assigns a risk category based on the customer’s aggregate score. Usually this type of model is constructed based on subject matter expert judgment or knowledge about the underlying process rather than formal analytical analysis. Because there is no specific underlying methodology or model design that firms must follow, they have an endless supply of model design and scoring options to choose from. However, this is also the weakness of these types of models, because the lack of a single statistical framework means there is no established statistical methodology for setting parameters or selecting variables to include in the model. Even after choosing a general modeling framework, firms must make numerous iterative adjustments to determine the combination of parameter settings that maximize the model’s fit to the target variable. In addition, there is no effective way to know whether the chosen parameter set is really the optimal set of values. This makes justifying these types of models to regulators more and more difficult.

Although heuristic, rule-based models were once the norm within the AML community because of their simplicity and ease of development, they are quickly being replaced by more scientific modeling approaches that can stand up to regulators’ scrutiny and that allow for a methodical approach to parameter setting and model validation.

STATISTICALLY BASED MODELS

Statistically based models are founded on well-established statistical methodologies and approaches that have been vetted, reviewed, and published in academic journals. Most statistically based models that financial firms use for customer risk rating are predictive models, such as linear regression, binary or ordinal logistic regression, decision trees (all types), and neural networks. The particular application and risk rating objectives determine the actual model that the firm selects. However, for customer risk rating, either binary or ordinal logistic regression models are currently the most common.

Unlike heuristic, rule-based models, statistically based models require that certain assumptions be met so that the modeling framework can be accurately tested and assessed to an acceptable degree of statistical confidence. A common goal in creating statistical models is to develop the simplest model—the one with the fewest variables—that is needed to make an accurate prediction. To select those variables, the firm must use a robust statistical framework based on widely accepted modeling approaches that maximize the likelihood that the target is estimated as accurately as possible. In addition, the firm can use standard approaches for assessing each variable’s significance to the model, gauging the model’s overall goodness-of-fit (to determine whether a more complex model is called for), and assessing the model’s predictive power—all of which it can also use to justify the model to regulators.

Although statistically based models have historically been less common in the AML community because they appear more complex and less understandable to the layperson, they are quickly becoming the industry standard in the face of the increasing regulatory pressure to use more scientific approaches. The ability to identify the variables that contribute most to the model, the selection of coefficients (that is, weights) based on maximum likelihood estimation, the ability to assess the strength of the model, and the ability to estimate the confidence of model predictions—all these advantages favor this modeling approach. The fact that regulators prefer these models only adds to the reasons to use them.

ORDINAL LOGISTIC REGRESSION

Once the firm has decided to move forward with a statistically based model, what type of statistical model should it select for its customer risk rating model? There are several methods to choose from, all with slightly different objectives and various strengths and weaknesses. But when the primary objective is to group customers into distinct buckets based on risk, ordinal logistic regression is a highly effective statistical modeling technique to consider. This is very often the situation a firm faces in developing a customer risk rating model, because the goal is to separate its customers into different risk groups based on their inherent risk characteristics.

Ordinal logistic regression differs from binary logistic regression in that the target variable can have more than two values and the corresponding categories are assumed to be ordered. Some logistic models allow for multiple categories of the target variable that aren't ordered (called multinomial), but ordered categories are preferred for two primary reasons:

- Ordinal models are simpler than multinomial models and therefore easier to interpret (Allison 2012).
- The hypothesis tests for ordinal models are more powerful than those for multinomial models (Allison 2012).

The cumulative logit model is the ordinal logistic regression model most commonly used; it is the default model used by SAS/STAT® software. The cumulative logit model assumes that the model can be combined into multiple binary splits of the dichotomous target variable, which is an assumption that must be initially tested using the score test for the proportional odds assumption. In cases where this assumption has been severely violated and cannot be reasonably believed to hold, usually a multinomial or binary model is used instead.

Although the cumulative logit model produces only one set of beta coefficients, the equation contains one less intercept constant than the number of target variables. The probability that an event will exist within each category is then calculated, and the category with the greatest probability is the one that the model assumes the customer belongs to (that is, this is the model’s estimated category).

The following sections walk you through the preliminary analysis that a firm must perform when it uses ordinal logistic regression to develop a customer risk rating model. By understanding the process at a high level, the firm can dispel the mystery and perceived complexity that surround these models. As a note, this paper assumes the use of SAS/STAT software; however, ordinal logistic regression is also available in other SAS® products, such as SAS® Enterprise Miner™.

DEVELOPING AN ACCURATE TARGET VARIABLE

Before you explore the data to be used in the model, you must evaluate the target variable for accuracy. The target variable for customer risk should reflect actual historical customer experience. If the firm is not confident in the accuracy of the risk being assigned to its customers by its current model, then it must sample and review customers across the different risk levels and attribute values. These samples must be large enough to be statistically significant for the building and testing of the ordinal logistic regression model.

MULTICOLLINEARITY

In general, multicollinearity occurs when two or more predictor variables (also known as independent variables or covariates) in a regression model are highly correlated with each other. To be more specific, multicollinearity exists when one or more of the variables used in the model can be linearly predicted within a reasonable degree of accuracy by using the other variables in the model. Note that this refers only to the relationship between the various predictive variables within the model; the predictive variables are expected to be correlated with the dependent, or target, variable.

When multicollinearity is present, the model's estimated coefficients can change erratically in response to small changes in the data or model. Although multicollinearity does not reduce the predictive power or reliability of the model as a whole, at least within the sample data used to train the model, it does affect calculations regarding individual predictions. That is, a regression model with correlated predictor variables can indicate how well all those variables predict the target variable, but it might not give valid results about any one predictor variable or about which variables are redundant.

There are two primary approaches to detecting multicollinearity within a model, and usually both are used. In the first approach, you can use the CORRELATION procedure in SAT/STAT to produce a correlation matrix to show the relationship between the predictive variables; it is important to use the Spearman correlation. The Spearman correlation coefficient is often described as being "nonparametric" and is used when data are grouped rather than numeric. The following statements illustrate the use of the CORRELATION procedure and the Spearman correlation:

```
proc corr data=YourData outs=CorrData(where=(UPCASE(_TYPE_)='CORR'))
     nomiss spearman;
  var YourVariables;
run;
```

In the second approach, you run a regression that includes all the predictive variables and request that the variance inflation factors (VIF) be provided. This produces a VIF value for each variable. The VIF is the reciprocal of one minus the coefficient of determination between the respective variable and the remaining predictor variables. Usually, a VIF greater than or equal to 4 indicates moderate multicollinearity, and a VIF greater than or equal to 10 signifies high multicollinearity. The following statements illustrate this approach by calling the REG procedure in SAS/STAT:

```
proc reg data=YourData;
  model yourvariables / TOL VIF;
run;
```

ZERO-COUNT CELLS

Zero-count cells, or events for which there are no observations, can cause unstable results within logistic regression. Furthermore, there is no maximum likelihood estimate for the respective variable. If this problem is not addressed, the SAS/STAT procedure issues warning messages that refer to either quasi-complete or complete separation of zero-count cells. However, the messages do not specify which variable the separation exists for, so it is important to investigate this further before you fit the logistic regression model.

In exploring the predictor variables, it is important to generate cross-tabulation tables of the individual predictor variables versus the target variable in order to identify zero-count cells. In the examples throughout this paper, the firm designates customer risk by using five categories: Low, Medium, High-Low, High-Medium, and High-High. This is to facilitate stratification across its high-risk customers.

Table 1 shows an example of quasi-complete separation, where there are no Low, Medium, or High-Low risk customers that also have a Country Risk of 3.

			Target Value (Risk)				
Variable	Data Type	Category	Low	Medium	High-Low	High-Medium	High-High
Country Risk	Ordered Category	1	102,528	117	268	68	31
" "		2	19,337	181	57	28	28
" "		3	0	0	0	33	73

Table 1. Example of Quasi-complete Separation

Table 2 shows an example of complete separation, where any customer with a Country Risk of 3 falls into the High-High customer risk category.

			Target Value (Risk)				
Variable	Data Type	Category	Low	Medium	High-Low	High-Medium	High-High
Country Risk	Ordered Category	1	102,528	117	268	68	31
" "		2	19,337	181	57	28	28
" "		3	0	0	0	0	106

Table 2. Example of Complete Separation

You can take the following actions to handle situations of quasi-complete and complete separation:

- Remove the variable or variables causing the problem (this is an option only if the variables contribute marginally to the model).
- Combine categories if the variable contains multiple categories.
- Define a rule outside the model that automatically sets customers to high-risk when they meet certain criteria that always result in those customers being considered high-risk.
- Check to see whether another variable is a dichotomous version of the variable in question.
- Grab more sample data that reflect what is missing, if possible.

PROPORTIONAL ODDS ASSUMPTION

The proportional odds assumption tests whether the coefficients of the dichotomous groupings of the outcome variable are the same. In ordinal logistic regression, the proportional odds assumption often does not hold. This fact is widely understood but often ignored because the practical implications, depending on the modeling objective, can be minimal. In SAS/STAT procedures, the score test for the proportional odds assumption tests the hypothesis that the estimated coefficients are not materially different from each other, regardless of the dichotomization.

Table 3 shows the mapped value of the target for the logistic regression model, and Table 4 shows the dichotomous groups that are used in the series of binary logistic regressions (that is, 1 versus 2, 3, 4, and 5).

Target	Model Value
Low	1
Medium	2
High-Low	3
High-Medium	4
High-High	5

Table 3. Example of Target Mapping

Dichotomous Groups	
0	1
1	2, 3, 4, 5
1, 2	3, 4, 5
1, 2, 3	4, 5
1, 2, 3, 4	5

Table 4. Example of Dichotomous Groups

The null hypothesis is that there is no statistical difference in the estimated coefficients between models. If the *p*-value is high, then the null hypothesis is not rejected, and you can conclude that the estimates are not significantly different. Output 1 shows an example of the score test results in SAS/STAT procedures. It is important to note that the score test often rejects the null hypothesis more frequently than it should. Stokes, Davis, and Koch (2012) state that this test needs at least five observations for each outcome of the category versus the target. In creating a cross-tabulation of a categorical variable versus the target variable, you need to have at least five observations in each cell. Otherwise, the sample size could be too small, there are simply no data (zero-cell problem), or it is a rare event.

Score Test for the Proportional Odds Assumption		
Chi-Square	DF	Pr > ChiSq
53.4698	24	0.0005

Output 1. Example of Score Test for Proportional Odds Assumption

MODEL DEVELOPMENT AND TESTING

In ordinal logistic regression, you must use a holdout sample to test the model to ensure that it truly fits the data and doesn't simply do so by chance, because many combinations of covariates are considered. A common practice is to build the model on approximately 70% of the data and test the model on the remaining data (the holdout data set). After testing, you can run the model on the whole data set. Then you can make comparisons between the outcome estimate percentages of the whole data set and those obtained during both model development and initial testing.

You can use the SURVEYSELECT procedure in SAS/STAT to create the build and test data sets. It contains a feature to randomly split the data. Or you can use the SAMPRATE= option to select a percentage at which to split the data.

The following statements show the SAMPRATE= option set to 0.7 (70%). The OUTALL option keeps all the records from the original data set; it creates a new variable called SELECTED that has a value of 1 if it was part of the 70% of the data and 0 if it was part of the remaining 30%.

```
proc surveyselect data=YourData out=SplitData samprate=0.7 outall;
run;
```

VARIABLE SELECTION METHOD

SAS/STAT procedures offer several different variable selection methods—forward selection, backward elimination, and stepwise selection—to help you determine which variables to include in the ordinal logistic regression model. However, your firm might choose to forgo these selection methods and instead use specific variables that it deems important with respect to the target variable—in this case customer risk.

In the *forward selection* method, the procedure systematically evaluates each available attribute and includes the effect in the model that adds the most to model performance. Then it goes through the remaining attributes one at a time to determine whether any others significantly improve performance. The procedure terminates when no further effects can be added to the model to significantly improve performance or when all the attributes are included.

In the *backward selection* method, the procedure begins with all the attributes. The variable that has the smallest partial *F* statistic is noted. The procedure systematically evaluates each available attribute and removes the effect in the model that is the most insignificant to model performance. Then it goes through the remaining attributes one at a time to determine whether any others are insignificant. The procedure terminates when the variable that has the smallest partial *F* statistic is significant.

In the *stepwise selection* method, the procedure systematically evaluates each available attribute and includes or removes the effect in the model that adds the most to performance. Then it goes through the attributes one at a time to determine whether their removal or addition significantly improves performance. The procedure terminates when no further effects can be added to or removed from the model to significantly improve performance or when all the attributes are included.

Although these selection methods work well in a mathematical sense, firms that are developing customer risk rating models should put additional thought into selecting variables that will also satisfy regulatory expectations. For instance, if variables for high-risk geography are not included in the model, then the firm should be prepared to explain why these variables were not significant (for example, all customers are located in high-risk areas). If it cannot find a good reason to exclude the variable, it should manually add the variable back into the model.

The following statements use the LOGISTIC procedure in SAS/STAT to build an ordinal logistic regression model with forward selection by using data from the build data set, and subsequently score the test data by using the built model:

```
proc logistic data=SplitData(where=(selected eq 1))
       plots(only)=(effect(polybar)
       oddsratio(range=clip)) descending outmodel=yourModel;
  class yourOrdinalVariables / param=reference;
  model risk= yourVariables / selection=forward rsq;
  output out=yourBuildResults predprobs=individual;
run;

proc logistic inmodel=yourModel;
  score data= SplitData(where=(selected eq 0)) plots)) fitstat
     out=yourTestResults;
run;
```

MODEL OUTPUT

Ordinal logistic regression calculates the odds ratios, coefficients, and other statistics for the significant predictor variables chosen by the model. It also calculates the probability of each risk level for a customer and assigns the risk level with the highest probability to that customer. You can use the risk assignments in turn to assess the predictive power. You can assess the predictive power of the model by using standard measures of association, which a SAS/STAT procedure can calculate for you. These predictive measures are derived from the concordant and discordant pairs observed within the data.

Figure 2 shows example estimates for the logit coefficients and odds ratios of the resulting intercepts and significant variables for an example ordinal logistic regression model. The coefficient indicates the change in the log odds ratio for each one-unit increase in the explanatory variable. For example, if a customer were to increase its SAR count by one, its risk score would be expected to increase 7.4422 units on the log odds scale, assuming that the other variables remained constant. The coefficients for the categorical variables are read differently, because the categorical values of the respective variables are compared to each other via “dummy” coding. For example, when you look at the comparison of “1 vs 3” for the variable Product Risk, you can expect a –8.693-unit decrease in customer risk as the customer moves from a product risk of 3 to 1.

The odds ratio is calculated by taking the exponent of the coefficient estimate for the respective variable. For example, the odds ratio estimate of 0.077972428 for the HIFCA Flag variable can be calculated by taking the exponent of –2.5514. This in turn can be read as the proportional odds of comparing a customer in a HIFCA area to a customer in a non-HIFCA area. In going from a non-HIFCA area to a HIFCA area, the odds of being in the highest risk level (5) are 0.078 times higher than the odds of being in the other combined lower risk levels (4, 3, 2, and 1), if all the other explanatory variables are held constant.

Variable Name	Comparison	Odds Ratio Est	Coefficient	Standard Error	Wald Chi-square	Pr > ChiSq
Intercept	5		1.4041	2.3805	0.3479	0.5553043
Intercept	4		5.1922	2.4435	4.5151	0.0335969
Intercept	3		9.8446	2.5708	14.6642	0.0001285
Intercept	2		14.9162	2.7964	28.4521	<.0001
Industry Code	1 vs 4	<0.001	-7.9945	2.4146	10.9621	0.0009299
Industry Code	2 vs 4	<0.001	-7.3831	2.4181	9.3226	0.0022634
Industry Code	3 vs 4	0.024192805	-3.7217	2.3299	2.5517	0.1101756
HIFCA Flag	0 vs 1	0.077972428	-2.5514	0.5566	21.0124	<.0001
CTR Count		>999.999	8.4971	1.7629	23.2307	<.0001
SAR Count		>999.999	7.4422	2.2739	10.7118	0.0010645
Product Risk	1 vs 3	<0.001	-8.693	1.1207	60.1694	<.0001
Product Risk	2 vs 3	0.027171137	-3.6056	0.5698	40.0372	<.0001
Expected Cash		4.176192724	1.4294	0.2097	46.4574	<.0001

Figure 2. Example Summary Table

Output 2 shows example estimates of the predictive power of the model (the values range from 0 to 1, with larger values signifying greater predictive power). The pseudo-coefficient of determination, often signified as R-square (R^2), is another popular statistic that you can use to assess the predictive power of the logistic model, where the Max Rescaled R-Square adjusts the statistic to account for the fact that in a discrete outcomes model the R-square value can often never actually equal 1.

The LOGISTIC Procedure

Probabilities modeled are cumulated over the lower Ordered Values.

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	96.9	Somers' D	0.947
Percent Discordant	2.2	Gamma	0.955
Percent Tied	0.9	Tau-a	0.708
Pairs	5705	c	0.973

Fit Statistics for SCORE Data			
R-Square	Max-Rescaled R-Square	AUC	Brier Score
0.758697	0.800536	.	0.308647

Output 2. Example Ordinal Logistic Regression Results

In general, tests of the model's predictive power assess how well you can predict the target variable by using the covariates (that is, predictive variables). It is entirely possible to have a model that predicts the target variable very well but fails the goodness-of-fit tests. It is also possible to have a model that makes poor predictions but shows very good model fit. Predictive power is commonly calculated using the measures of association of Somers' D, gamma, tau-*a*, and c (or AUC), which are all listed in Output 2.

Another useful way to view the model results is to generate a two-way contingency (cross-tabulation) table of the predicted target versus the actual target, as shown in Figure 3.

Model Error Severity (Combined Data)						
Estimated Target	Actual Target					Row Total
	Low	Medium	High-Low	High-Medium	High-High	
Low	57	1	1			59
Medium	2	47	4			53
High-Low		1	32	13	2	48
High-Medium			2	18	3	23
High-High				3	25	28
Total	59	49	39	34	30	211

Field Key
Correct Prediction
Inaccurate by 1
Inaccurate by 2
Inaccurate by 3
Inaccurate by 4

Error Rates	Count	Percent
Correct Prediction	179	84.83%
Inaccurate by 1	29	13.74%
Inaccurate by 2	3	1.42%
Inaccurate by 3	0	0.00%
Inaccurate by 4	0	0.00%
Total	**211**	**100.00%**

Figure 3. Example Contingency Table

You can also use the LOGISTIC procedure to produce an output data set that contains the predicted customer risk along with the predicted probabilities for each level of risk. This is very useful because firms usually want a score that is associated with the risk level assigned to each customer. A score can be calculated by using the sum of the weighted probabilities.

If you have five risk levels (1–5), the score would fall between 1 and 5. However, you can apply a scale by multiplying the weighted probability of the score by some factor. In Figure 4, Customer X would be assigned a risk level of 5 (High-High) because it has the highest calculated probability. To calculate the weighted probability, you add the weighted probabilities together. Because you have five risk groups, multiplying the sum of the weighted probabilities by 20 results in scores from 20 to 100. Figure 4 shows an example of these numbers.

	Customer X				
	Low	Medium	High-Low	High-Medium	High-High
Probability	0.000000000	0.000000038	0.000004454	0.000943457	0.999052050
Weight	1	2	3	4	5
Weight * Probability	0.000000000	0.000000077	0.000013361	0.003773830	4.995260252
Weighted Probability	4.99904752				
Scale to 100 Points	99.98095039				

Figure 4. Example of Scoring

DEPLOYMENT

During the deployment step, the model logic is implemented in the operational production system. There are many ways to do this, including SAS web services, batch SAS processing, various queuing servers, and recoding the model logic into the language that the operational production system expects. For this process, it is assumed that the model will be deployed as a SAS batch process and that input data delivered by the firm match the data delivered for the modeling process.

MODEL RISK GOVERNANCE

When the customer risk rating model is created and put into operation, the firm must create a plan for ongoing model validation, as described in *Supervisory Guidance on Model Risk Management* (OCC 2011-12). Deliverables might include the validation of the target variable used to train the model, the validation of the model performance, and a model validation report.

Firms might want to perform analysis to assess the relationship between the target variable (Customer Risk Rating) and the actual resulting number of scenario alerts generated, case referrals, or Suspicious Activity Reports (SARs) filed on the customer. This enables the firm to document that customers receiving a Customer Risk Rating of 5 (High-High) are more likely to be involved in suspicious activity than customers that receive a rating of 4, 3, 2, or 1. If it is determined that the target variable lacks the desired degree of accuracy, then the firm has the option to update the target variable setting and retrain the model on the revised data set. This would provide a model that more accurately identifies customers that meet the firm's definition of "risky."

Regulators expect all analytical models to be validated on an ongoing basis, as described in OCC 2011-12. This process involves tasks such as periodically assessing of the model's performance, ensuring that appropriate model controls are in place, determining that the right covariates are included in the model, adjusting the coefficients as needed, and so on. It can also include testing new variables that were not available before or that contained only sparse data.

Firms should also produce a model validation report every year that documents *all* model validation tests that it has performed and the results of those tests. The report should also contain the following:

- a description of the model, its parameters, its input variables, and its strength and weaknesses
- validation of all model components, including input data, assumptions, processing, and reports
- an evaluation of the model's ongoing conceptual soundness, including development details that might be relevant
- evidence of ongoing monitoring, including process verification and benchmarking
- an outcomes analysis, including back-testing

CONCLUSION

In addressing customer risk rating by financial firms, SAS has found that statistically based models are the most effective method of classifying customer risk while satisfying the regulatory expectation that quantitative modeling techniques and practices be used. Although ordinal logistic regression modeling requires a variety of analyses and testing, this approach has been accepted in the Anti-Money Laundering community as a successful statistically based method of measuring customer risk. The key to any AML modeling is to ensure that development, testing, and validation are well documented and explainable to regulators. Increased regulatory pressure requires continuous evaluation of a firm's customer risk. To meet these demands, the AML community has turned to analytical and statistical methodologies to improve customer risk assessment in order to effectively manage risk and identify the customers that require the most attention and review.

REFERENCES

Allison, P. D. (2012). *Logistic Regression Using SAS: Theory and Application*. 2nd ed. Cary, NC: SAS Institute Inc.

Stokes, M. E., Davis, C. S., and Koch, G. G. (2012). *Categorical Data Analysis Using SAS*. 3rd ed. Cary, NC: SAS Institute Inc.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors:

Edwin Rivera
SAS Institute Inc.
Edwin.Rivera@sas.com

Jim West
SAS Institute Inc.
Jim.West@sas.com

Carl Suplee
SAS Institute Inc.
Carl.Suplee@sas.com