WHITE PAPER

Developing Scenario Segmentation and Anomaly Detection Models

How analytics can be used for AML compliance programs

sas
THE POWER TO KNOW.

Contents

Contributors

Jim West, Principal Analytical Consultant for Fraud and Compliance, SAS

Carl Suplee, Director of Product Management, Security Intelligence Practice, SAS

Introduction

Since 2012, financial institutions have increasingly adopted more rigorous analytics to improve their anti-money laundering monitoring programs.[1] And although institutions are motivated to find new ways to optimize AML transaction monitoring processes, it's not clear if their reasons for performing such improvements have changed over time.

Firms are quickly finding themselves under oversight scrutiny. Because of enhanced regulatory pressure to continuously evaluate the firm's risks, identify emerging trends, report suspicious activity and expediently make changes, firms are seeking out new and aggressive approaches. To meet these demands, the AML industry has turned to analytical/statistical methodologies to improve monitoring programs by reducing false-positive alerts, increasing monitoring coverage, and reducing the rapidly escalating financial cost of maintaining an AML program.

Already struggling to control costs, firms are continuously scrutinizing the economic cost to perform AML compliance. AML officers are judged not only on their ability to react to regulatory changes and quickly implement solutions, but also their accountability for AML program expenses. AML officers are increasingly required to play multiple roles, and it takes real leadership to balance and manage all of these expectations.

There are two real questions: Are firms learning how to effectively blend quantitative and qualitative transaction monitoring approaches in order to implement a risk-based program? Or are they solely relying on one or the other? Banks often rely solely on one strategy instead of effectively blending both methods.

Segmentation Is the Logical First Step

Then where do banks start? An effective AML transaction monitoring strategy begins with a sound foundation for monitoring customer activities — and a quality segmentation model provides just that foundation. Banks can begin with segmenting the customer base by analyzing customer activity and risk characteristics.

Segmentation is the process of grouping customers and accounts that have similar characteristics and transactional behaviors, with the objective of setting risk-based thresholds that are appropriate for each particular segment. Segments of customers and/or accounts may be grouped together based on one or more of their inherent characteristics, such as:

- Average transaction amount
- Average transactional volume
- Net worth
- Product usage
- Region
- Line of business

[1] In 2012, the US Office of the Comptroller of the Currency (OCC) began applying the OCC's *Supervisory Guidance on Model Risk Management* (OCC 2011–12), including SR 11–7 and DEF Regulation Part 504, to anti-money *laundering compliance practices.*

- Customer risk-rating classification
- Industry type, or
- Customer/account type

Segmentation is the primary foundation for risk-based scenario threshold setting, and the quality of the segmentation model directly affects the transaction monitoring system's ability to perform in an effective and efficient manner.

A quality model that groups homogeneous groups of customers and/or accounts – and allows separate monitoring for high-risk groups of customers and/or accounts – provides the opportunity to set appropriate threshold levels for monitored segments. It also allows for enhanced monitoring of high-risk segments. In addition, a quality segmentation model helps you set scenario threshold values to provide effective coverage throughout the customer, account and external entity populations. In fact, most banks that perform transaction monitoring without the use of a segmentation model (e.g., separating only personal and commercial customers) find that they have very poor alert coverage within their customer and account populations. As a result, they generate the vast majority of alerts for the largest and lowest-risk customer groups, and few alerts are generated for smaller, high-risk customer groups.

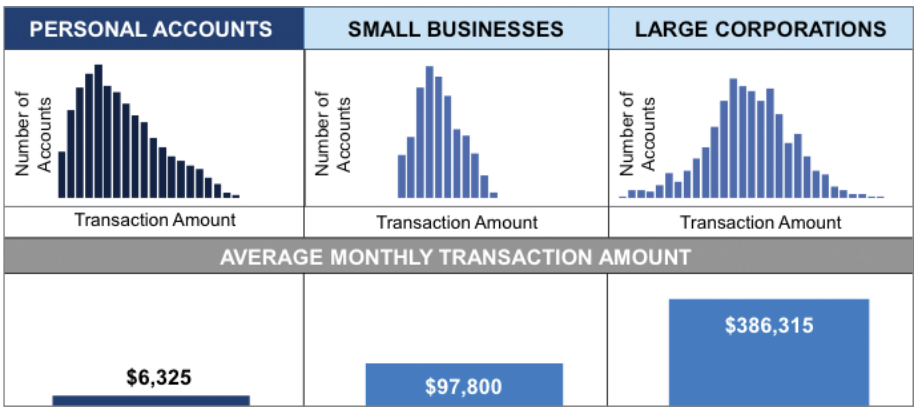

Figure 1: Segmentation analysis of monthly activity by customer type.

The SAS® Approach

The SAS approach to segmentation generally requires three primary activities:

1. Customer, account or external entity population segmentation (or a combination thereof).
2. Further refinement of individual segments into peer groups (only needed if anomaly detection will be performed).
3. Initial threshold setting (needed to assign the scenario threshold parameter values to use initially prior to the first scenario tuning and model verification project).

In addition, SAS adheres to the guiding principles of OCC 2011-12 when developing, implementing and validating segmentation and peer group models, including the process of initial threshold setting.

Development of the Segmentation Model

The development of the segmentation model is further broken down into 11 subtasks. They include:

Model framework and analytics data development

1. Model framework development decision process.
2. Segmentation data attribute availability, mapping and reconciliation.
3. Develop the segmentation analytics data file (combining data from various sources needed for the analysis, as well as the creation of data summary fields).

Data quality and exploration

4. Perform the data quality and completeness assessment (i.e., verify that the data being used for the segmentation model development is materially free from defects).
5. Conduct the categorical attribute exploration.
6. Conduct the transactional data exploration.
7. Develop the customer profile summaries (summaries of the various customer groupings identified).

Segmentation model development and verification

8. Perform the segmentation model framework exploration (generally several different models are considered using different attributes).
9. Develop the segmentation model based on the results of the segmentation model framework exploration.
10. Verify the developed segmentation model to ensure that it meets the organization's needs and contains homogenous groupings in terms of risk and activity.
11. Develop the comprehensive segmentation report.

Before the actual segmentation work can begin, you should undertake the segmentation model development decision process, where you consider the goals, objectives and scope of the segmentation model. In addition, you should decide on the focus of the segmentation model — be it at the customer, account or external entity level — depending on your bank's needs.

While SAS party-focused scenarios generally require a customer-level segmentation model, you can run account-focused scenarios using either a customer- or account-level model. In the case where you run an account-focused scenario using a customer-level segmentation model, the segment containing the primary customer on the account is used for determining the account's applicable segment for alert generation.

Although developing separate models for account- and party-focused scenarios allows for more refined threshold setting, it also means that you must develop, maintain and periodically validate two models — all adding additional cost and complexity. And your bank may also need an external entity segmentation model if you are monitoring customers at the external entity level (common in correspondent banking). This is also

a good time for you to identity the model dimensions to be considered. Commonly, segmentation models have either two or three dimensions, such as a risk dimension, activity level dimension, and a line of business or product type dimension.

The first step in developing the segmentation model is to identify the data attribute availability, field mapping and reconciliation of the customer population. It's imperative that your model developer fully understands your bank's data, including the customer and account types, the various roles on the transaction, and the different types of transactions conducted by your bank. This is when your model developer would perform a reconciliation of the customer population so that everyone is clear on which data records are being included in the model development, and which are being intentionally excluded. This is also when you would want to identify potential risk attributes (data points) to consider when segmenting, including:

- Client type.
- Industry codes (e.g., NAICs, SICs).
- Income.
- Revenue.
- Account or product types.
- Geography/location.
- Transaction types.
- Channels (e.g., branch, ATM, online).
- Transaction volumes and/or amounts.
- Customer risk ratings.
- Other customer attributes.

The key to proper risk attribute selection is to conduct data discovery and clearly understand the data lineage, quality and sourcing. While numerous attributes may exist that potentially could be pertinent to the segmentation model development, poor data quality and inconsistent collection may prevent their use. As the regulatory guidance explains, it's important to specifically call out any weaknesses or limitations in the data that will be used to analyze the customer population and develop the segmentation model.

Bringing the various data sources together into a single file that contains the needed information is often more difficult and time consuming than is initially expected. For example, you need to summarize account-related information at the customer level when building a customer-focused model. This requires the use of numerous "dummy" fields that identify the various aspects of the entities you're analyzing.

Before using the collected data to begin building the segmentation model, understand that the end product will only be as good as the data going into it. Explore your bank's data and generate some basic metrics to look for, such as: missing values, extreme values (outliers), inappropriate values, customers without accounts, and accounts without transactions or with excessive transactions. The results of this study will help your bank ensure that the data you use to build the model is free from material defects. It's also a good practice to review the data summaries as a preliminary test to see

whether they seem reasonable given your bank's knowledge of the business. Also keep in mind that the regulations specifically mention a model data quality assessment.

Both the categorical and transactional data exploration involves identifying the customer groups and associated characteristics that your bank feels add superfluous AML risk to your business. There are common entity types and geographical regions that the FFIEC manual and FinCEN have listed as potentially affecting AML risk. However, the true extent to which these risk factors affect a bank's AML risk will depend on the customer base, key products and internal controls. It's also important that you properly assess the transactional variability within the population, since greater variability will often require a greater number of activity-based segments contained within the final model. This is where determining both the wire and cash utilization percentages can come in handy to help understand the nature of the customer activity. At this stage, you can use various analytical/statistical approaches to explore the data and determine the relationships between the variables, and identify key attributes to segment similar customers and accounts together.

You can rely on graphical approaches to dissect the population under review and understand the relationship between the various customer and account attributes. The advantage of using graphical techniques is that they allow your analyst to quickly get a general sense of the population distribution, correlation between variables and the variability (or spread) within the data. These approaches can include scatter plots, frequency plots, histograms, box plots, stacked bar charts, pie charts, heat maps and so on. Graphical plots are also an effective way to identify outlying groups of customers and/or accounts. SAS offers several high-performance products for developing graphical visualizations, including SAS Visual Analytics and SAS Visual Statistics.

In addition to the use of graphical approaches, common statistical methods and tests can further refine the various groups within the population. Various clustering approaches help to identify groups of entities with similar characteristics. While k-means is probably the most commonly used clustering technique within AML, there are certainly more advanced clustering algorithms that you can apply to provide your bank with superior ways to identify separation within the population.

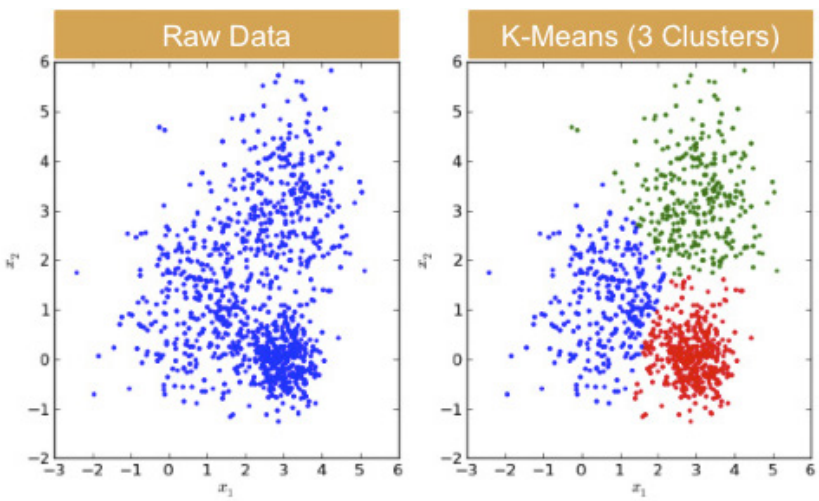

Figure 2: K-means clustering technique.

All the information gathered above allows your model developer to create customer profiles that identify groups of customers that indicate similarities and lists out products used, transaction amounts/volumes and account types held. The customer profiles are often of great interest to banks, as they provide a better understanding of the customer population the banks are monitoring.

To begin development of the segmentation model framework, you should start with the business knowledge of the customer population and historical alert experience. Then further refine the various model framework options based on the analysis performed above.

After you determine the key data attributes and identify high-risk customer groups, it's time to develop the various initial segmentation model frameworks (or model designs.) The primary goal here is simply to form a high-level segmentation model framework based on the findings above. Normally, your bank will have a number of different model frameworks (or model designs) to decide from, each focusing on slightly different customer attributes and risk factors. It's important that your bank consider various model frameworks so you can demonstrate to regulators that your bank was open to different models, and that you selected the final model design after weighing the pros and cons of each.

Once you've selected the segmentation model framework that's best for your bank's business, it's time to further refine the model. During this phase, you can split a single segment containing large variability into several smaller, more homogeneous groups; you can combine different segments found not to be overly heterogeneous; or you can add more segments to accommodate customers containing similar previously unconsidered risk attributes. The main goal is to adjust the model as needed to best meet your bank's initially stated objectives.

It's important to keep in mind your bank's ultimate goal in building the segmentation model: to allow for effective risk-based transaction monitoring by applying different threshold values to different groups of customers. Analysts, who commonly have a strong academic background, often get bogged down in trying to decide the optimal segmentation or clustering approach to use rather than focusing on the true goal: to allow effective risk-based thresholds to be set and promote alert coverage within the customer/account population.

Before spending lots of time implementing complex clustering algorithms or developing advanced quantitative segment assessments, analysts should ask themselves this question: "Will using this advanced approach allow for significantly greater risk-based transaction monitoring, and will the regulators and management understand what I've done?" Keep in mind a fundamental rule in model building: All else being equal, a simpler model is preferred over a more complex model.

Now that your bank has done its analytical homework, it's time to evaluate the data used to support the final segments within the model. There may still be a need to separate segments into smaller, more homogeneous segments. There are three considerations:

1. Does the segment have a wide range of total monthly transactional activity?
2. Does the segment still comprise several independent statistical distributions?
3. Does the segment contain customers with widely varying degrees of inherent AML risk?

As part of the final model's verification process, perform statistical tests in order to:

1. Verify that segments of customers are in fact from different subpopulations.
2. Identify outliers.
3. Validate normality of the population (or non-normality, which is more often the case).

It's also good to generate distributional metrics to be used in determining the similarity of the segments, including the skewness, kurtosis, coefficient of variation, mean, median and the empirical percentile distribution.

Once your bank has completed its analysis and developed the segmentation model that best fits your firm's needs, there's no better time to put together comprehensive documentation (in accordance with the OCC's 2011–12 Supervisory Guidance on Model Risk Management). The value of clearly documenting the entire segmentation process and providing a concise overview cannot be overstated. In the minds of many regulators, if it wasn't documented, it wasn't performed.

Do not overlook the effective challenge discussion in the OCC 2011–12 guidance. Take the time to ensure your bank has picked the best segments based on what is known (and has considered alternatives during the model framework exploration process). But understand that this will not be the last time that your bank will evaluate its segmentation model. It needs to continuously monitor and conduct segmentation analytics to ensure that the model is still valid and that the underlying customer population or product offering hasn't changed, causing a deterioration in the segmentation model over time. A segmentation model is a "model" as defined by the OCC and thus requires ongoing periodic validation just like any other model used by your bank.

Development of the Peer Group Model

The primary reason you develop a peer group model is to allow for anomaly detection at the peer group level. This allows you to monitor anomalous activity against both the individual customer's historical transactional activity and against the customer's peers (i.e., other customers that behave in a similar manner and/or contain a similar risk profile). Your ultimate goal of monitoring activity at the peer group level is to identify customers who are expected to behave in a similar fashion as their peers, but don't.

Peer groups are essentially more refined segments, and thus the peer group model is developed after the bank approves the segmentation model. Each of the segments has its customer population broken down into subsets based on the customer attribute(s) selected for differentiating between the various customers (note that different segments can use different attributes when defining peer groups). The attribute(s) can be used to separate the segment into peer groups by rank order, classification or clustering techniques, depending on the data type(s) involved and the goal of the particular peer group model.

When developing peer groups, it's important to balance the need for homogeneity within the groups with the number of customers contained in each group. Peer groups should never comprise less than 100 customers since this can result in having a statistically invalid population size from which to compare the individual customers against. The most common customer attributes considered for peer group development are:

- Transaction amount (average or total).
- Transaction count (average or total).
- Customer net worth (i.e., the sum of the account balance where the customer is listed as the primary).
- Industry type code (for commercial enterprises).
- Geography (in combination with another attribute).
- Transaction type (generally used in combination with either transaction amount or count).

However for all practical purposes, peer groups can be defined based on any set of customer attributes that allow similar customers to be grouped together.

For the rank-order allocation approach, the attribute for each customer is sorted from smallest to largest. Then the customers are separated into N different equally sized groups based on the ordered attribute values. This approach is generally used when the attribute used to create the peer group involves the transaction amount, transaction count or net worth amount.

With the classification approach, various distinct groups of attribute categories are created and customers are then assigned to peer groups that contain an attribute value within the group. This approach is generally used when the attribute used to create the peer group involves the geography, industry code or a unique combination of attributes.

The cluster analysis allocation approach utilizes a clustering technique, such as k-means, to assign the customers to various peer groups based on one or more customer attributes. While it is possible to use this approach when considering only one attribute, it is more commonly used when multiple attributes are used. This method also makes creating relatively equally sized peer groups difficult to do and is heavily affected by outliers contained in the data.

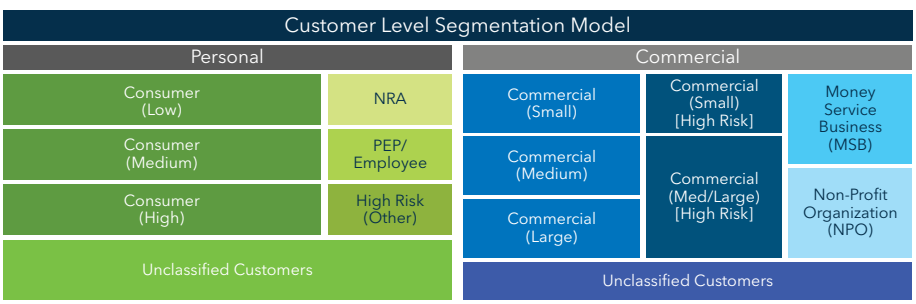

Figure 3: In the customer level segmentation model, customers are assigned to various peer groups based on one or more customer attributes.

Conclusion

Enhanced regulatory pressure requires continuous evaluation of your bank's risks. To meet these demands, the AML industry has turned to analytical/statistical methodologies to:

- Improve monitoring programs.
- Reduce false-positive alerts.
- Increase monitoring coverage.
- Reduce the rapidly escalating financial cost of maintaining AML programs.

A well-designed segmentation model can significantly increase your bank's AML monitoring coverage of the customer population while focusing the alert generation through risk-based thresholds on the customers that pose the most AML risk. This involves generating a greater proportion of alerts compared to the underlying population for some higher-risk segments, and a lower proportion of alerts as compared to the underlying population for less risky segments. For example, a bank that historically generated 80 percent of its personal banking alerts solely on the top 10 percent of customers (based on average monthly transactional amount) may find that after segmentation, the same personal banking customer population is generating only 5 percent of alerts. The other 75 percent of the alerts that previously were being generated for those customers are now alerting on higher risk customers and customers well below the 90th percentile in terms of average monthly transactional amount.

Effective peer grouping and anomaly detection can enhance existing AML monitoring programs and find customer outliers that would not normally be picked up by traditional heuristic pattern-based rules. This will augment the AML program by identifying false negatives or unusual activity that would otherwise go unnoticed.

By effectively blending both quantitative and qualitative methods, banks can monitor more effectively by segmenting the customer base and tuning the scenarios to identify the activity that poses the most risk to the bank.

For more information on high-performance products for developing graphical visualizations, visit sas.com/va and sas.com/visualstatistics.

To contact your local SAS office, please visit: sas.com/offices