Paper 2675-2018

Risk Pathways: Using Machine Learning Techniques of SAS® Viya™ to Understand Customer Risk Drivers

Leigh Ann Herhold, Zencos Consulting LLC

ABSTRACT

With the new SAS® Visual Analytics and SAS® Visual Data Mining and Machine Learning capabilities on SAS® Viya®, institutions can more quickly and easily understand drivers of customer risk in a visually appealing way. Analysts can identify new opportunities for stakeholders to identify potential gains or losses by understanding drivers of risk and detecting new risk factors as they emerge over time. Using a combination of supervised and unsupervised machine learning methods, businesses can further understand and improve their own definition of customer risk as an organization. In this paper, Zencos shares its expertise in identifying high-risk attributes, understanding the relative importance of each driver, and recognizing key combinations of factors associated with risky behavior through an anti-money laundering case study.

INTRODUCTION

Identifying and mitigating customer, industry and company risk is a continuous and evident challenge across all industries. Organizations lacking an effective and robust risk management strategy are oftentimes penalized with fines and negative news from regulatory agencies. Figure 1 shows a small collage of recent negative news resulting from regulatory fines. In 2017, approximately $19 million[1] was issued for HIPAA violations, $128 million[2] for BSA-AML civil money penalties, and $23.7 million[3] for Consumer Product Safety Commission (CPSC) penalties. On a broader scope, banks have globally paid $321 billion in fines since 2008 from regulatory penalties related to money laundering, market manipulation, and terrorist financing (Cox, 2017). Although regulatory agencies such as the Office of the Comptroller of the Currency (OCC) provide guidelines to model risk management (OCC 2011-12), organizations often struggle to develop defendable methodologies that can be maintained and tuned over time that properly represent customer risk as risk evolves. This paper outlines a data-driven approach to help identify, understand, and validate customer risk by using SAS Visual Analytics and SAS Viya's machine learning and analytical tools in an AML case study.

Figure 1. Collage of Recent Newspaper Headings of Regulatory Fines

[1] Figure estimated from Compliancy Group HIPAA Done Right: https://compliancy-group.com/hipaa-fines-directory-year/

[2] Figure estimated from BankersOnline: https://www.bankersonline.com/penalty/penalty-type/bsa-aml-civil-money-penalties

[3] Figure estimated from Statista: https://www.statista.com/statistics/619556/civil-penalties-issued-by-us-cpsc-since-2006/

COMMON RISK MANAGEMENT CHALLENGES

Risk is having exposure or opportunity for gain or loss, but is oftentimes associated with a negative connotation. It is important to be able to quantify and gauge risk in comparison to other human activity, which is somewhat unpredictable and uncontrolled. Because risk is essentially a perception, there is a flexibility in judging risk and a number associated with risk is meaningless without being used as a point of comparison. There are several business challenges in developing a robust risk management strategy, and this next section will discuss three different notable challenges associated with AML risk.

NEW RISKS EMERGE

Risk is not a static condition. With more available data and the constant digital and technological transformations to keep up with customer demand, new risks are constantly emerging. This creates additional pressure for the Compliance groups within financial institutions to have a flexible array of tools for risk management, even as new risks emerge over time.

According to Accenture's 2017 Compliance Risk Study, compliance executives are "anticipating an 89% increase in compliance investment over the next two years" for banks, insurers, and capital market firms (Accenture, 2017). When it comes to new risk emergence, it can be difficult for Compliance groups to move from a reactive risk management approach to a proactive approach.

This gives organizations the ability to stay on top of managing risk as it evolves over time. Companies can fall back into a reactive approach when they are using outdated and defective models, or are mis-interpreting models. A way to counter this is to have an established methodology to identify new risks and tune existing models. The Risk Pathways methodology addresses both new risk emergence and tuning.

IDENTIFYING RISKY BEHAVIOR WITH ACCURACY

Identifying truly risky behavior with accuracy in a timely manner is difficult. Approximately 5% of all investigative efforts for evaluating money laundering risk are productive – the remaining 95% are false positives. It's important to use business expertise to primarily identify initial high-risk factors but also to leverage data and algorithms to understand how combinations of factors contribute to the risk classification of a customer. Machine learning algorithms can be used to detect unknown patterns, relationships, and outliers within the data that can lead to more accurate classification of customer risk. Later this paper will highlight how to visually understand how combinations of attributes contribute to a customer's classification of risk.

INVESTIGATIVE EFFICIENCY

Maintaining efficiency throughout investigative efforts can be difficult, as EDD, alert and case reviews can take a few minutes up to many hours. This becomes a labor-intensive feat especially without the right tools to aid the investigative process. Keeping cost at bay while conducting thorough investigations is a top concern for many financial corporations. This challenge dovetails with identifying truly risky behavior. By properly identifying customer risk, investigative teams can better prioritize investigative efforts towards alerts and cases that are more likely to be productive. Proper scenario tuning and threshold setting can reduce the backlog of excessive alerts that have a low probability of being productive.

According to ACAMS's paper on AML Rule Tuning, "in order to maintain AML detection scenarios current with best market practices, rule tuning exercises have become an ever-increasing important task to perform at a financial institution" and the "tuning needs to be constant" (ACAMS, 2015). Having an automated process, clear dashboards, intuitive tools, effective machine learning algorithms, and a streamlined approach will greatly improve investigative efficiency.

RISK PATHWAY METHODOLOGY

Risk Pathways is a data-driven and analytical methodology that combines both machine learning techniques using SAS VA and SAS Viya with business knowledge to create an alternative approach to supplement either new or established customer risk management strategies. This methodology is demonstrated through an Anti-money laundering (AML) case study focused on identifying each customer's opportunity to launder money through a financial institution. This methodology is to be utilized to help meet regulatory and compliance standards, aid an institution in understanding the main drivers of risk, and validate the risk ranking process. Before diving into the analytical process, it's important to understand the regulatory requirements around which this process is designed.

AML COMPLIANCE AND REGULATORY REQUIREMENTS

According to the UNODC, Global money laundering transactions are estimated to be 2% to 5% of global GDP, or roughly $800 billion - $2 trillion U.S. dollars annually (UNODC, 2018). Section 312 of the USA Patriot Act requires U.S. financial institutions to perform due diligence and enhanced due diligence (EDD) for customers identified to have a high opportunity or risk of laundering money or terrorist financing. Enhanced due diligence is an additional review that involves higher scrutiny of the individual. The FFIEC outlines several considerations for conducting an EDD, and oftentimes this review including assessing the purpose of the account, source of funds, political exposure, negative news, banking references, account activity and changes, and related parties. Productive EDD reviews resulting in confirmation of high risk status oftentimes result in the filing of a Suspicious Activity Report (SAR). Figure 2 shows the common due diligence steps incorporated into an AML monitoring system.

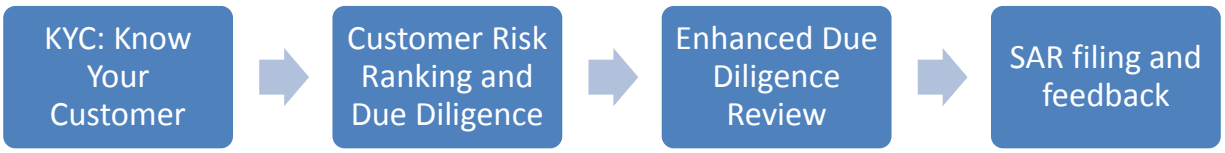


Figure 2. Common Due Diligence Process for AML Investigations

HIGH-LEVEL ANALYTICAL PROCESS

Risk Pathways provides a way to risk rank customers, prioritize investigative efforts, and use a data-driven risk management approach that will allow organizations to identify where investigators are making decisions inconsistently and therefore validate business processes. In the OCC 2011-12, the OCC and Federal Reserve provide guidance for model risk management and pay attention to model validation, so this case study will focus the validation process. The high-level components of this case study include a two-layered modeling approach and are shown in Figure 3.

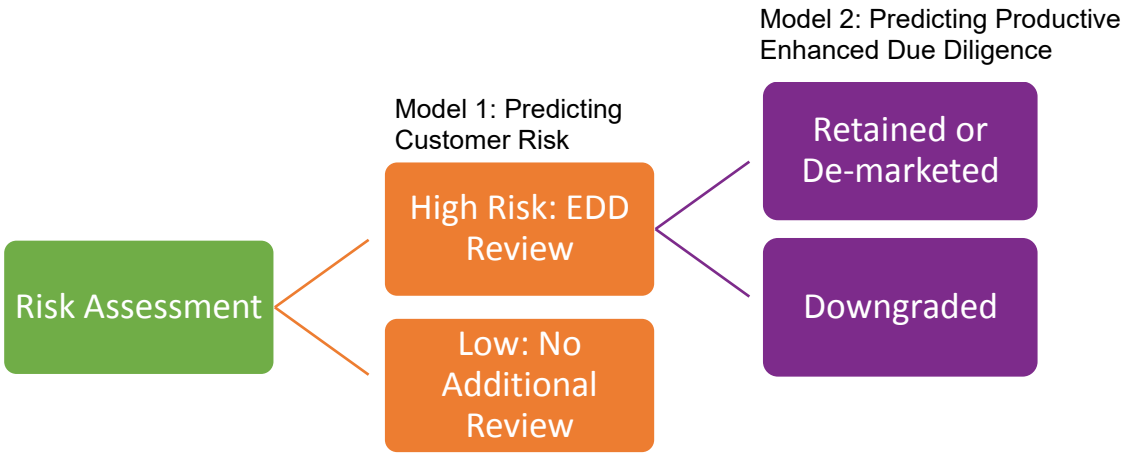
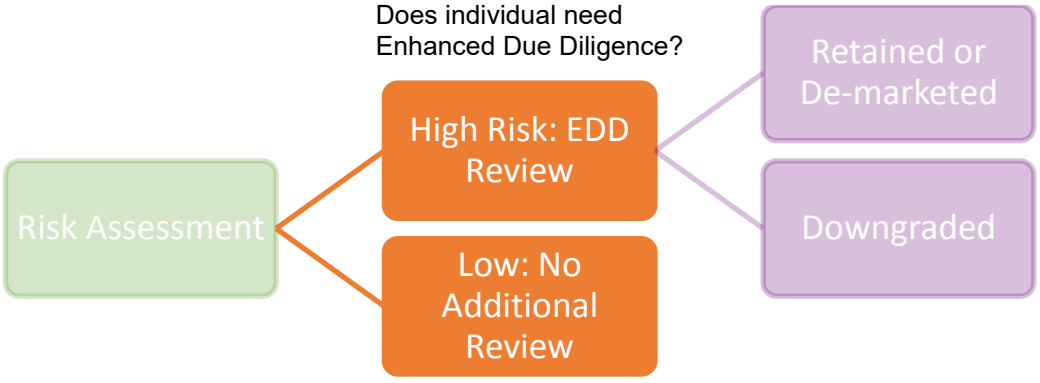


Figure 3. High-Level Process Flow

An initial risk assessment is performed based on a business-level risk assessment, highlighted in Figure 4.

Figure 4. Initial Risk Assessment

Analytical models are applied to transactional and attribute data to provide insight into whether an individual has a high probability of needing an additional review. Customers are identified by analysts as either low risk and not needing additional scrutiny or high risk and requiring an enhanced due diligence review.



After the EDD review, customers are then either retained at a high-risk status and in some cases, de-marketed, or their risk ranking is downgraded, highlighted in Figure 5. Analysts and investigators document reasoning of why an individual's risk status was updated or maintained, oftentimes providing a rich source of (unstructured) text data. Here there is another opportunity to use analytical modeling to predict the updated risk status, given that an individual was selected for an EDD review.

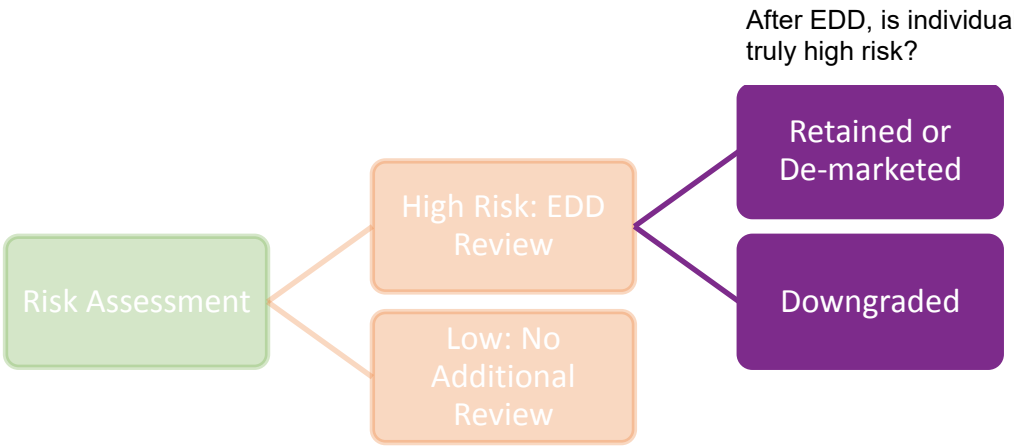


Figure 5. Post-EDD Disposition

The next section will walk through the technical components of this two-step modeling process by using graphical features offered in SAS VA for data exploration and SAS's Data Mining and Machine Learning capabilities on SAS Viya for predictive analysis.

DOWN IN THE DATA: DATA EXPLORATION FOR MODEL BUILDING

SAS VA is a point-and-click interface that is easy to use for data exploration and understanding the customer population. This case study was conducted using a sample of credit union transactional data for approximately 3,300 customers. Each customer within this sample has been determined by the business as either being at a high or low risk based on their perceived opportunity to launder money. Approximately 10% of the sample was determined as high risk and the remaining 90% as low risk, as shown in the hollow pie graph in Figure 6.

Figure 6. Pie Graph of Customer Sample by Initial Risk Disposition

Several of the main drivers in determining that initial risk classification include the following:

- Transaction Country
- Residence Country
- Transactional Activity (ACH, Cash, Wire)
- Politically Exposed Person (PEP) indicator
- Historical Suspicious Activity Reports (SARs)

It is important for analysts to use both business knowledge as well as other data attributes that are not necessarily used for the initial risk assessment in the initial model. Combinations of attributes can oftentimes contribute to model accuracy.

Data Analysts and Data Scientists oftentimes look at frequency distributions of activity to understand common behavior. Within this sample, most customers do not appear to be heavy cash users. This observation is validated by the heavy right-tailed distribution shown in Figure 7 that represents a frequency distribution of cash transactions within the last three to six months.

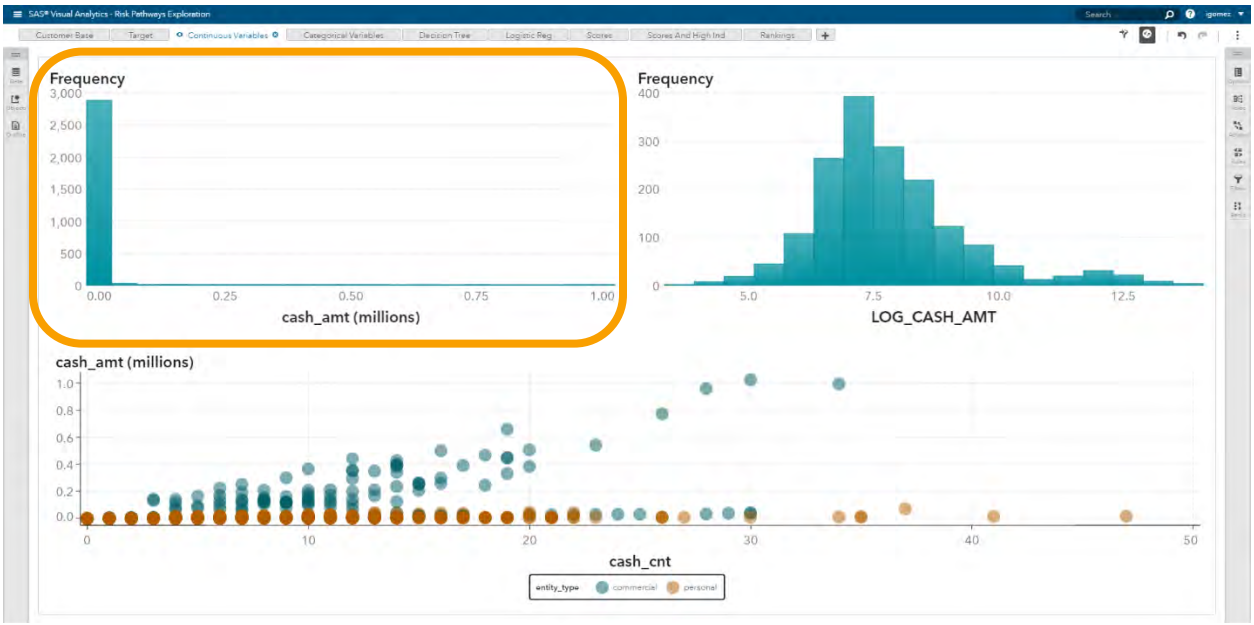


Figure 7. Right Tailed Distribution of Cash Amount in Millions

Analysts can also use scatterplots to understand relationships between variables. The relationship between number of cash transactions and the amount transacted between both personal and commercial individuals is plotted in Figure 8 below. The scatterplot provides evidence for the fact that this sample of commercial customers appear to send or receive cash in higher amounts. If the sample is representative of the population, then these inferences can also be applied to the population.

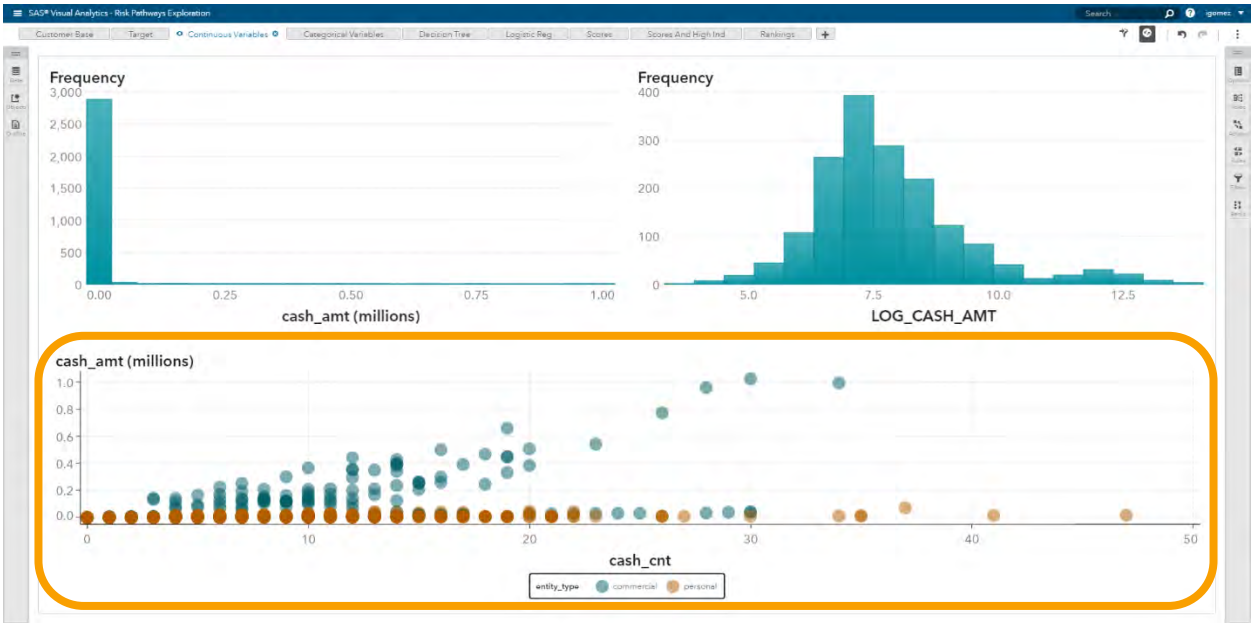


Figure 8. Scatterplot of Cash Amount and Cash Count by Entity Type

Businesses highlighted in green that reside in the upper right hand corner of the scatterplot in Figure 9 not only have more cash transactions, but also have a high frequency of cash transactions. Perhaps these businesses are cash-only businesses such as nail salons, small restaurants, food trucks, or laundromats. Analysts can combine useful pieces of transactional activity such as cash frequency and cash amount to create ratios, correlations, and other statistics to incorporate into a predictive model.

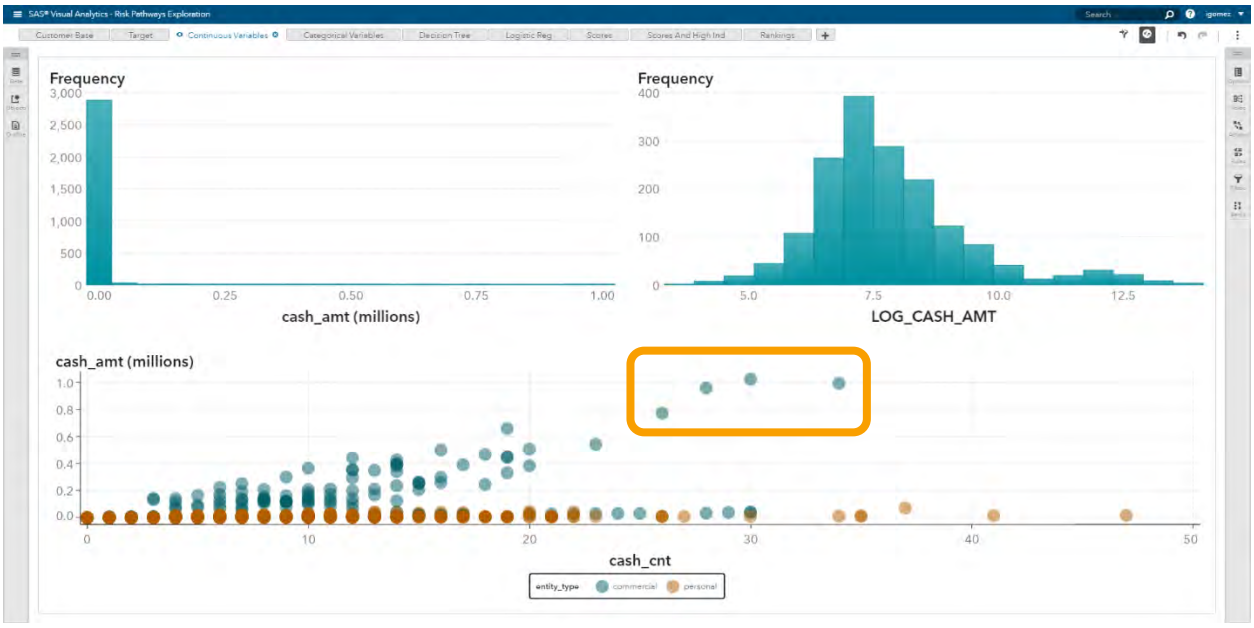


Figure 9. Scatterplot Highlighting Predominately Cash Businesses

CREATING THE BASELINE MODEL

SAS Visual Data Mining and Machine Learning on Viya provides several predictive modeling and machine learning model options including logistic regression, linear regression, neural networks, decision trees, etc. For the purposes of creating a baseline risk classification using a target variable of high or low risk, a decision tree or logistic regression model is appropriate. Figure 10 shows an example of a decision tree created in SAS VA on Viya with interactive leaves and branches that will display basic statistics. Analysts can iteratively tune and grow the tree using SAS VA's interactive mode.

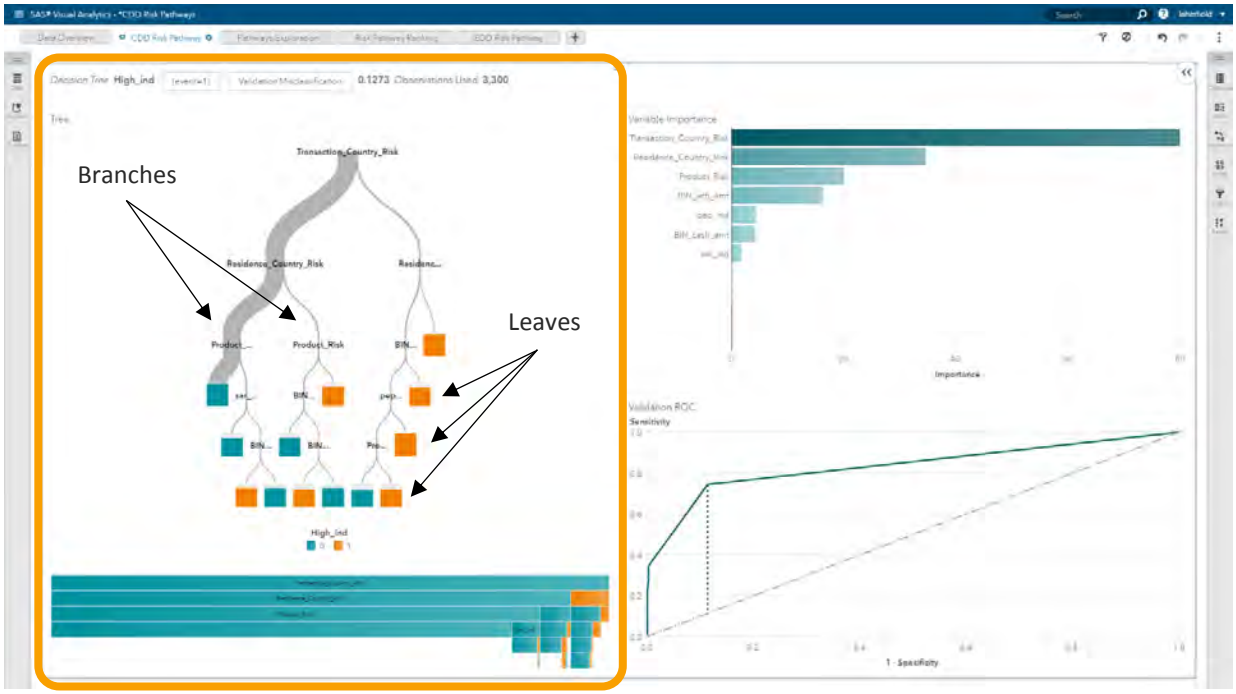


Figure 10. Decision Tree for CDD Baseline Model

Decision trees can be easily explored with the visual and interactive display that SAS VA provides. The root node at the top of the tree corresponds to the best predictor. In this case study, the country in which the transaction took place had the greatest predictive power in determining an individual's risk classification. The branches, or gray lines, flow from the top of the graphic to the bottom and end with either a green or orange leaf. The thickness of the branch indicates the relative number of customers. There are different customer attributes at each split in the tree. These variable or attribute splits are mathematically determined to partition the data into subsets with homogenous values.

Each leaf indicates a bucket of customers, and the color of the leaf indicates whether the majority of the customers are predicted by the model to be high or low risk – buckets of high risk customers are orange and buckets of low risk customers are green. In addition to the color-coding, the size of the leaf is also proportional to the number of customers with a that color-coded outcome.

The decision tree node in SAS Visual Analytics also automatically provides a variable importance plot and a few different options for validation curves, both shown on the right-hand side of Figure 11. The validation curves provide model accuracy statistics. The next section will discuss the baseline model findings and validation of business assumptions.

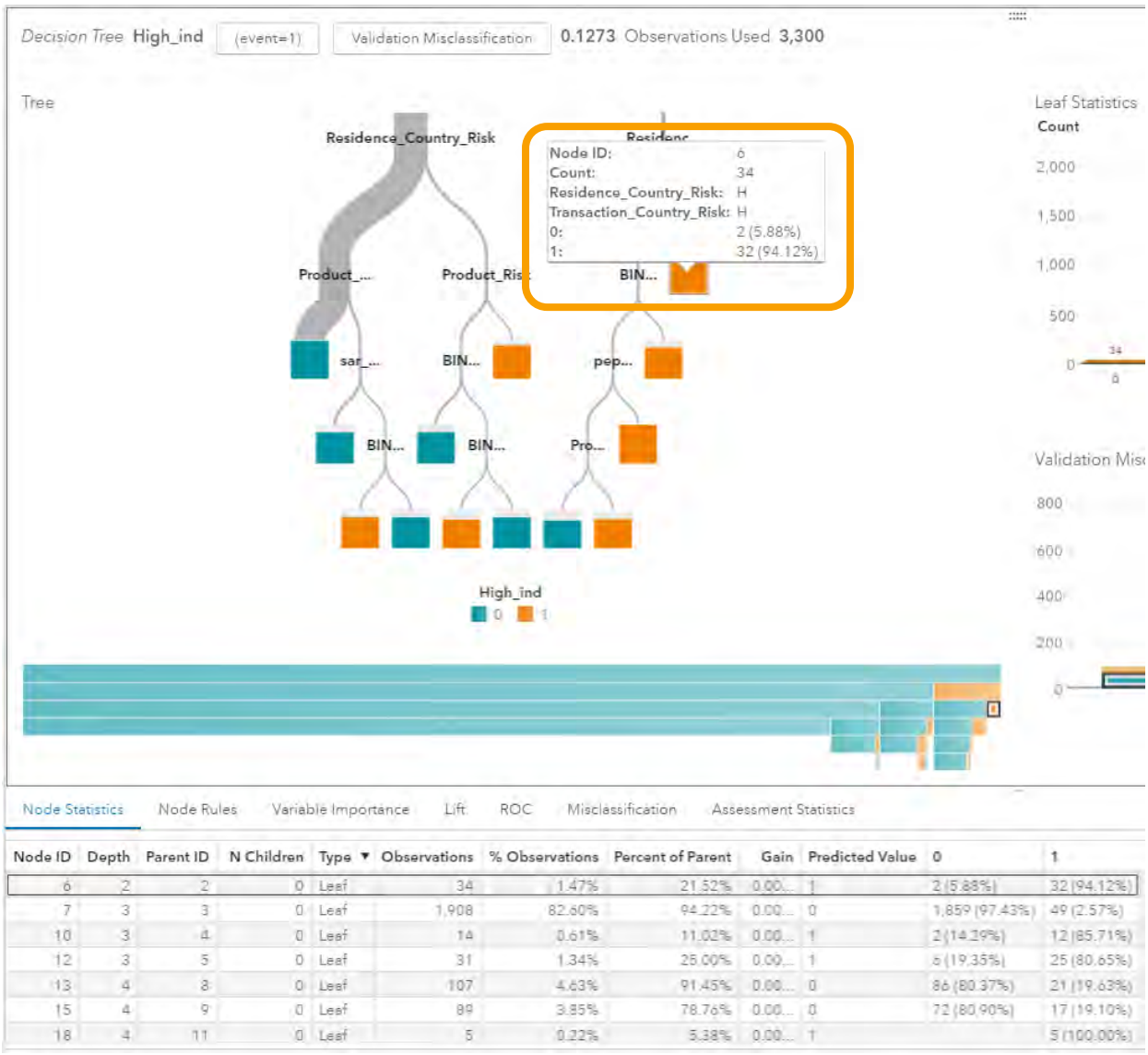
Figure 11. Decision Tree Variable Importance Plot and Validation Curve

RISK VALIDATION AND MODEL INSIGHTS WITH SAS VA ON VIYA

What conclusions can we make from this baseline model? By visualizing the risk pathways, ie. the combinations of attributes that determine customer risk, we can validate and re-affirm business assumptions about the customers within the population. In addition, the model results allow for interactive examination of risk attributes, their relative importance, and provide an opportunity to re-address any inconsistencies in the initial risk assessment of individuals with similar attributes.

Each leaf is a group of customers with common attributes. Most customers within this sample have transactional activity in low risk countries, a low residence country risk, and low product risk, which ultimately classifies them as low risk customers. There are several other pathways or combinations of attributes that lead customers to be identified by the model and being high risk.

Figure 12 shows an example of one node and its statistics. In this group, there were 34 individuals that were all predicted by the model as being high risk. However, 2 of the 34 were determined by the business as being low risk; the remaining 32 were determined by the business as high risk. This discrepancy reveals an inconsistency in how the initial risk was determined by the business, which can be readjusted or addressed if necessary. It is also possible that additional information used to make initial risk assessments for these customers is not included in the baseline model and should be added.



Node Statistics | Node Rules | Variable Importance | Lift | ROC | Misclassification | Assessment Statistics

Node ID	Depth	Parent ID	N Children	Type	Observations	% Observations	Percent of Parent	Gain	Predicted Value	0	1
6	2	2	0	Leaf	34	1.47%	21.52%	0.00...	1	2 (5.88%)	32 (94.12%)
7	3	3	0	Leaf	1,908	82.60%	94.22%	0.00...	0	1,859 (97.43%)	49 (2.57%)
10	3	4	0	Leaf	14	0.61%	11.02%	0.00...	1	2 (14.29%)	12 (85.71%)
12	3	5	0	Leaf	31	1.34%	25.00%	0.00...	1	6 (19.35%)	25 (80.65%)
13	4	8	0	Leaf	107	4.63%	91.45%	0.00...	0	86 (80.37%)	21 (19.63%)
15	4	9	0	Leaf	89	3.85%	78.76%	0.00...	0	72 (80.90%)	17 (19.10%)
18	4	11	0	Leaf	5	0.22%	5.38%	0.00...	1		5 (100.00%)

Figure 12. Example of Node Statistics within Decision Tree

RISK PATHWAY RANKING

An additional step to help prioritize investigative efforts is to rank the risk pathways in a meaningful way in context of the business problem. A variety of node statistics are available within SAS VA that can be used to develop the ranking methodology. One example of how to rank the pathways is by using a weighted percentage of high risk individuals per pathway. Table 1 shows each leaf node and the corresponding percentage of customers that the business defined as either low risk (Business Defined 0) or high risk (Business Defined 1) and also provides the model prediction in terms of 1 (High Risk) or 0 (Low Risk). Here the risk pathway classification is simply determined as either a high-risk or low-risk pathway. Of course, this pathway ranking classification can become much more granular.

One interesting finding from this ranking is that although Leaf 7 is a low risk pathway and contains the majority of the low-risk customers within the sample, it also contains the highest number of customers determined to be high risk by the business. This would require additional review and possibly additional risk attributes into the model.

Node ID	Type	Risk Pathway Ranking	Business Defined 0	Business Defined 1	Predicted Value	Risk Pathway Classification
18	Leaf	1		5 (100.00%)	1	H
6	Leaf	2	2 (5.88%)	32 (94.12%)	1	H
12	Leaf	3	6 (19.35%)	25 (80.65%)	1	H
10	Leaf	4	2 (14.29%)	12 (85.71%)	1	H
21	Leaf	5	5 (26.32%)	14 (73.68%)	1	H
24	Leaf	6	2 (25.00%)	6 (75.00%)	1	H
19	Leaf	7	1 (20.00%)	4 (80.00%)	1	H
7	Leaf	8	1,859 (97.43%)	49 (2.57%)	0	L
23	Leaf	9	57 (71.25%)	23 (28.75%)	0	L
13	Leaf	10	86 (80.37%)	21 (19.63%)	0	L
15	Leaf	11	72 (80.90%)	17 (19.10%)	0	L
20	Leaf	12	4 (80.00%)	1 (20.00%)	0	L
22	Leaf	13	4 (80.00%)	1 (20.00%)	0	L

Table 1. Risk Pathway Ranking by Node

MODEL TUNING & MACHINE LEARNING: PREDICTING CUSTOMERS FOR ENHANCED DUE DILIGENCE

Customers predicted to be high risk by the baseline model are candidates for requiring an additional enhanced due diligence review. After creating a baseline model, a secondary model can be used to predict and understand the combinations of attributes that contribute to an individual having a productive enhanced due diligence review. For this case study, a productive EDD review results in maintaining the customer status at high risk or de-marketing the individual. Figure 13 shows the results of a secondary decision tree developed to predict whether an individual is likely to have a productive Enhanced Due Diligence (EDD) review.

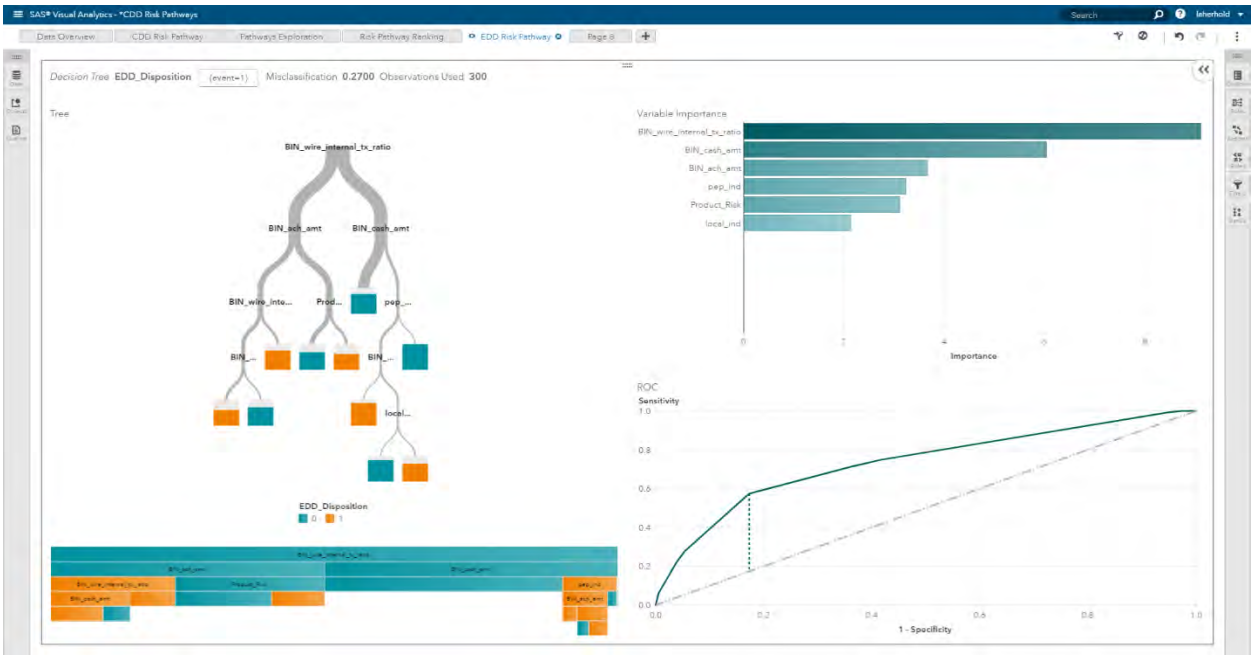


Figure 13. Secondary Decision Tree Model for Predicting Customers Likely to Need an EDD Review

During EDD reviews, analysts or investigators maintain a running log of notes to document decisions and findings. Machine learning techniques such as topic discovery or text mining can be applied to historical EDD review notes to provide additional insight into other key attributes that investigators are looking at that could impact their decision. New topics or themes extracted from text mining can be incorporated into variables and added as candidates for the original model, given that the data is easily gathered or calculated for the entire population. This provides a great way to continually tune the original baseline model in addition to the secondary model as well as capture or identify new risks that investigators are naturally examining during their EDD reviews. This creates a feedback loop as shown in Figure 14.

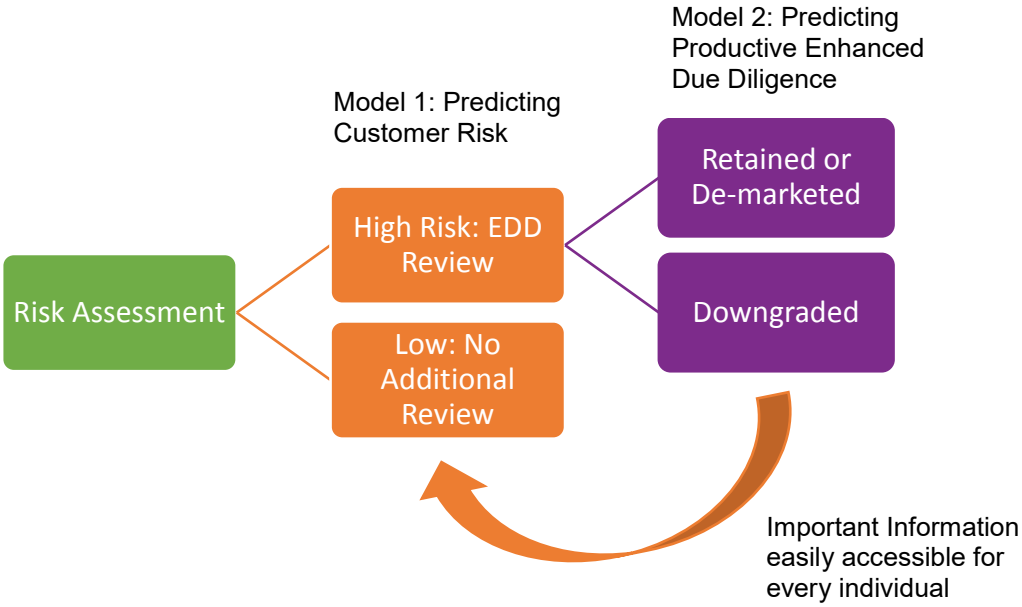


Figure 14. Feedback Loop within Process Flow for Model Tuning

In addition to looking at overall model performance statistics, risk pathway rankings can be compared both before and after tuning or having incorporated new information extracted from text mining of EDD notes. This comparison can highlight improvements and changes in the model before and after tuning. Figure 15 shows an example of comparative node statistics for the original and tuned models. One caveat with this approach is that customers removed from one leaf or pathway will be placed into another. Multiple comparisons of pathways are necessary to validate this type of tuning.

Original CDD Model Low Risk Pathway

Node ID	Type	Business Defined 0	Business Defined 1	Predicted Value	Risk Pathways Classification
13	Leaf	86 (80.37%)	21 (19.63%)	0	L

Tuned CDD Model Low Risk Pathway

Node ID	Type	Business Defined 0	Business Defined 1	Predicted Value	Risk Pathways Classification
13	Leaf	78 (83.87%)	15 (16.13%)	0	L

Figure 15. Comparison of Original vs. Tuned CDD Model Results for One Risk Pathway

ENHANCEMENTS

There are several enhancements to this methodology that are worth mentioning and should be considered in a production environment, including model ensembling, customer segmentation, neural networks, and scorecard risk ranking.

MODEL ENSEMBLING

Model ensembling is a method of combining or "ensembling" multiple models together to enhance the predictive power and accuracy of the overall outcome. This enhancement may be much more feasible for larger financial institutions who have teams of analysts developing models to improve risk accuracy. This includes random forest modeling and bootstrap aggregation, which is a machine learning algorithm used to improve the stability and accuracy of machine learning algorithms as well as reduce variables and avoid model overfitting (inaccurately inflating predictive model accuracy).

CUSTOMER SEGMENTATION

Customer segmentation is a common and valuable unsupervised method used to identifying sub-groups within a population and identifying characteristics of how an institution's population differs. Perhaps the business would use the customer segments to treat groups differently when assessing relative risk. This can also be used as a technique for outlier analysis.

SCORECARD RISK RANKING

Creating a scorecard and calculating each customer's individual risk score would provide a more granular prioritization, especially for helping investigators address the riskiest customers first. Scorecards are usually built from logistic regression models – a modeling technique also provided in SAS Viya – and are easily understood by both analysts and business groups.

DEEP NEURAL NETWORKS

With the new capabilities on SAS Viya 3.3 and upcoming versions, more types of neural networks are available on Viya. Neural Networks can be applied in a variety of ways – to predict customer risk, to more accurately identify outliers, object recognition, etc. Although neural networks are not easily explainable, they are able to detect more complex patterns that go undetected by other models.

CONCLUSION

Although regulatory agencies provide guidelines and suggestions for risk classification, model risk management, and model validation, there is a lot of flexibility in each institution's approach to quantify, classify, and defend an individual's risk. As new risks emerge over time, compliance groups are pressured to have controls in place to be able to easily adapt customer risk assessments in a mathematically defendable way to meet regulatory requirements. Risk Pathways provides one example of an approach that can be incorporated into an organization's current methodology for defining and validating customer risk.

REFERENCES

Accenture. (2017). 2017 Compliance Risk Study: Financial Services. Retrieved March 06, 2018 from https://www.accenture.com/us-en/insight-compliance-risk-study-2017-financial-services

Compliancy Group LLC. (2018). HIPAA Fines Listed by Year. Retrieved March 06, 2018 from https://compliancy-group.com/hipaa-fines-directory-year/

Cox, J. (2017). Now that Banks Have Paid $321 Billion In Fines, Here's the Toughest Test Ahead. Retrieved March 06, 2018 from https://www.cnbc.com/2017/03/03/BANKS-HAVE-PAID-321-BILLION-IN-FINES-SINCE-THE-CRISIS.HTML

Lucchetti, U. Jr. (2015). AML Rule Tuning: Applying Statistical and Risk-Based Approach to Achieve Higher Alert Efficiency. ACAMS. Retrieved March 06, 2018 from http://www.acams.org/wp-content/uploads/2015/08/AML-Rule-Tuning-Applying-Statistical-Risk-Based-Approach-to-Achieve-Higher-Alert-Efficiency-U-Luccehtti.pdf

UNODC. (2018). Money-Laundering and Globalization. Retrieved March 06, 2018 from https://www.unodc.org/unodc/en/money-laundering/globalization.html

ACKNOWLEDGMENTS

Thank you to Eric Hale, Director of Fraud and Compliance Solutions at Zencos Consulting, for his contribution in jointly developing out the risk pathways methodology and sharing his AML expertise.

RECOMMENDED READING

- SAS Global Forum Paper 875-2017 Optimizing Anti-Money Laundering Transaction Monitoring Systems Using SAS Analytical Tools by Leigh Ann Herhold, Stephen Overton, and Eric Hale; Zencos Consulting. Accessed at: http://support.sas.com/resources/papers/proceedings17/0875-2017.pdf
- SAS Visual Data Mining and Machine Learning Documentation at https://support.sas.com/documentation/prod-p/vdmml/index.html
- FFIEC Manual: Customer Due Diligence – Overview https://www.ffiec.gov/bsa_aml_infobase/pages_manual/olm_013.htm
- OCC 2011-12: Supervisory Guidance on Model Risk Management https://www.occ.treas.gov/news-issuances/bulletins/2011/bulletin-2011-12a.pdf

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Leigh Ann Herhold
Zencos Consulting
(919) 237-9079
laherhold@zencos.com