Paper 6643-2016

Making Better Decisions About Risk Classification Using Decision Trees in SAS Visual Analytics©

Stephen Overton and Ben Murphy, Zencos Consulting

ABSTRACT

SAS Visual Analytics© Explorer puts the robust power of decision trees at your fingertips for an opportunity to visualize and explore how data is structured. Decision trees help analysts better understand discrete relationships within data by visually showing how combinations of variables lead to a target indicator. This paper explores the practical use of decision trees in SAS Visual Analytics Explorer through an example of risk classification in the financial services industry. This paper will explain various parameters and implications, explore ways the decision tree provides value, and provide alternative methods to understand the reality of imperfect data.

INTRODUCTION

Decision trees are powerful tools for visualizing relationships within discrete and continuous data. They use a combination of nodes and branches displayed in a tree fashion to differentiate how variables can predict values of a response variable. Figure 1 provides a sample output of a decision tree. Each branch of a decision tree splits values into different bins based on the relationships which exist within the data as well as how the decision tree is configured in the exploration. Each node in the decision tree provides the analyst a breakdown of how the response variable is distributed at each leaf of the tree.

This paper provides guidance for building decision trees in SAS Visual Analytics Explorer. Readers of this paper should have a basic understanding of SAS Visual Analytics Explorer and how to prepare data for analysis in SAS Visual Analytics. The second half of this paper illustrates how decision trees are beneficial by showing how customer segmentation and profiling within the BSA/AML industry can be approached using decision trees.

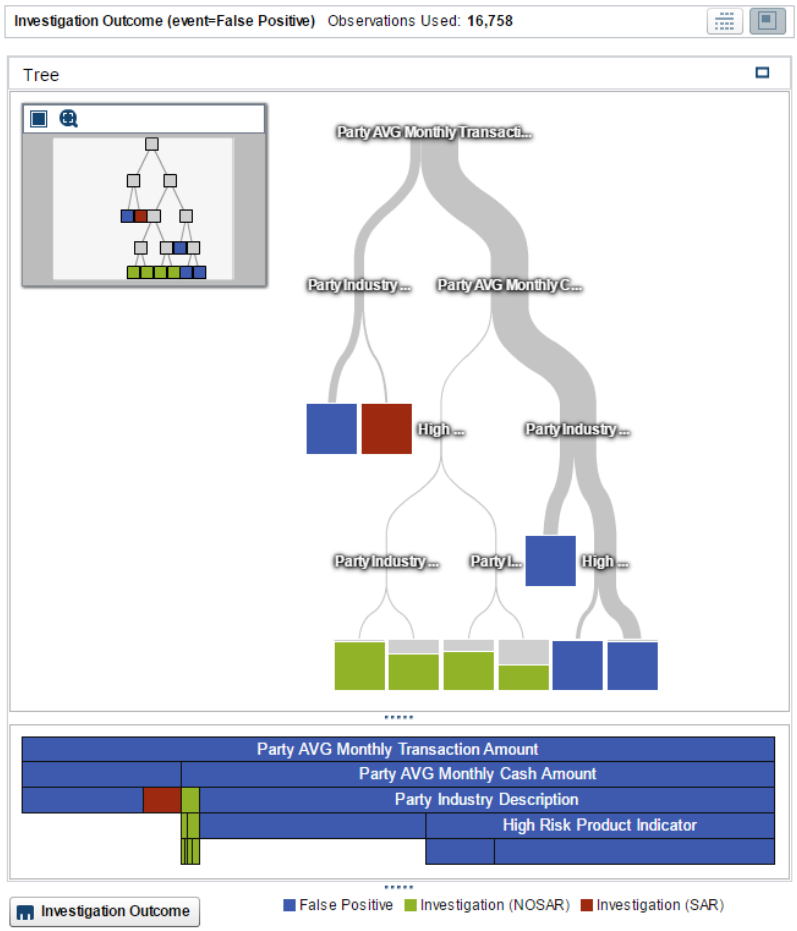

Figure 1: Sample Decision Tree output from SAS Visual Analytics Explorer.

Note: This paper is based on SAS Visual Analytics version 7.3 with SAS Visual Statistics licensed. SAS Visual Statistics enables additional advanced features for decision trees and should be considered when applying knowledge from this paper.

HOW TO BUILD A DECISION TREE IN SAS VISUAL ANALYTICS

Decision trees are helpful tools for business analysts to use when they understand the business data and want to try to discover patterns. With a robust statistical methodology behind an intuitive and user-friendly interface, analysts can leverage decision trees to explore relationships between data elements.

In SAS Visual Analytics, decision trees can be built within a data exploration by assigning one variable to the response role, and then assigning one or more variables to the predictor role. This is shown in Figure 2. The response variable should be the variable that the analyst is trying to understand more deeply. In the example outlined later in this paper, the goal is to identify patterns in productive alerts, which lead to more productive investigations using the Investigation Outcome as the response variable. The response variable may also be considered the dependent variable in other analytical methodologies. The analyst will also assign one or more variables as predictors, and the SAS Visual Analytics engine will consider all of the selected variables as potential predictor variables when trying to build a decision tree to understand the response variable.

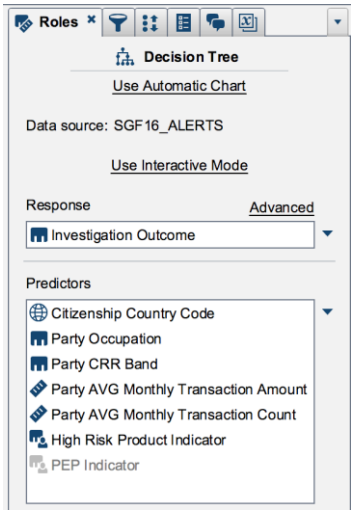

Figure 2. Decision Tree Roles.

The first step in getting SAS Visual Analytics to help build the decision tree is to add the response variable. As soon as predictor variables are added, SAS Visual Analytics will try to build the best tree possible after each new predictor variable is added. SAS Visual Analytics will add levels and branches to the tree showing the predictor variable used at each level. The number of levels and branches that are created depends on the settings and parameters that are specified on the Properties tab in the upper right.

The default parameters are typically useful for most analysts, but the power of SAS is at your fingertips as SAS Visual Analytics allows you to customize several aspects of how the decision trees are built. As always, the SAS Visual Analytics User Guide provides tremendous detail and help on every detail.

ADVANCED PROPERTIES FOR DECISION TREES

At the start, most users will probably want to see how each of the Growth Strategy options performs for their data. The Basic and Advanced use default parameters that make for a maximum of six levels in both settings and a maximum of two branches and four branches per level, respectively. If desired, users can always set the Growth Strategy to custom, which allows the user to change maximum branches and maximum levels to any other values, as well as customize the leaf size, predictor bins, and scale the pruning level. The Properties tab is shown in Figure 3.

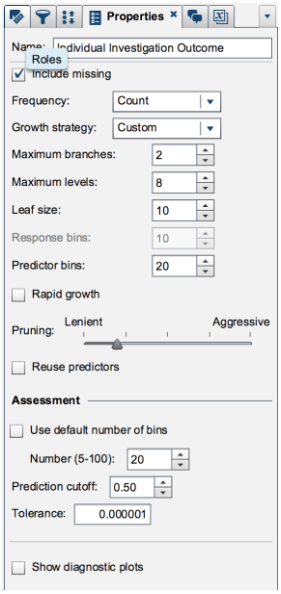

Figure 3. Advanced Properties for Decision Trees.

Once a basic decision tree is up, the properties tab can show diagnostic plots that can be useful in determining how well the tree is predicting different outcomes. Start with the leaf statistics by node id, which shows how often a leaf node identifies one specific outcome, and similarly the misclassification bar chart will show how often the tree correctly predicts each of the responses.

If the diagnostic plots show that the decision tree is not performing well, the Custom growth strategy also allows users to customize many other settings. If the predictor role variable is not discrete, the predictor bins parameter will setup bins for the continuous variable. The leaf size setting will require a minimum number of observations in each of the leaf nodes to ensure that the decision tree is not over fitting observations into leaf nodes in small quantities. The Rapid Growth option will utilize different methods for building the tree and generally the default value of unchecked is suitable for most users, but statistically minded users that want to use k-means fast search can check the rapid growth box. The Custom growth strategy allows users to reuse predictor variables multiple times in case that would produce a more accurate decision tree.

The Custom growth strategy also provides a sliding scale to adjust the pruning level. The scale ranges from lenient to aggressive and the value selected on the pruning scale will determine how strong the explanatory power of branches and leaf nodes must be to remain on the decision tree. More aggressive pruning will require a higher explanatory power and more lenient pruning will retain more leaves and branches with lower predictive accuracy.

UNDERSTANDING RISK USING DECISION TREES IN SAS VISUAL ANALYTICS

As a part of BSA/AML compliance (Bank Secrecy Act/Anti-Money Laundering), customer segmentation and profiling is an important exercise banks and financial institutions must perform on an ongoing basis as a part of evaluating and identifying risk. Decision trees give AML analysts the ability to breakdown what attributes of a customer differentiate productive versus non-productive alerts. Decision trees also allow AML analysts the ability to understand transactional behavior to profile spending habits of customer segments. These exercises only scratch the surface of building an effective AML model tuning program but serve as great examples to show the power of decision trees within SAS Visual Analytics.

IDENTIFYING RISKY CUSTOMER SEGMENTS WITHIN PRODUCTIVE AML ALERTS

To demonstrate customer segmentation modeling using decision trees, let's first define the business context and assumptions around the data. The data used in this example represents AML alerts triggered from an automated transaction monitoring system. The purpose of each record is to indicate potential suspicious activity for a bank's customer at a given point in time, based on a range of scenarios designed to monitor for money laundering and terrorist financing activities. Transaction monitoring systems typically produce a few hundred to a few thousand alerts each run date, which can run daily, weekly, or monthly, depending on the financial institution's monitoring needs.

Key variables used in this demonstration are described in Table 1 below.

Variable	Focal Entity	Description
Alert ID	Alert (grain of the table which metrics are based upon)	Unique identifier for the alert.
Alert Create Date	Alert (grain of the table which metrics are based upon)	When the alert was created. This variable can be used as a filter in the decision tree to set specific periods of time if needed.
Investigation Outcome	Alert (grain of the table which metrics are based upon)	Response variable used in decision tree which indicates the outcome of the investigation. Values in this field tell us if the alert was a false positive or if the alert was productive.
High Risk Product Indicator	Party (customer associated with alert)	Boolean value which signifies if the customer performs transactions using known high risk products such as prepaid cards, trust accounts, or electronic banking.
MSB Indicator	Party (customer associated with alert)	Boolean value which signifies if the customer the alert fired for is a Money Service Business, which is a known high risk entity.
Non Profit Indicator	Party (customer associated with alert)	Boolean value which signifies if the customer the alert fired for is a non-profit organization, which is a known high risk entity.
Party CRR Band	Party (customer associated with alert)	Classification of party (customer) into risk categories such as HIGH, MEDIUM, and LOW. These are a part of the financial institutions Customer Due Diligence (CDD) program.
Party Industry Description	Party (customer associated with alert)	NAICS code description of the party such as GAS STATION, REAL ESTATE BROKERAGE, or CASINO. This field only applies to organizations or businesses (party_type_desc).
Party Occupation	Party (customer associated with alert)	Occupation of party such as ATTORNEY, JEWELER, or MILITARY. This variable only applies to individuals or persons.
Party Type Desc	Party (customer associated with alert)	Signifies if the party is a person or individual, or an organization such as a business or large corporation.

Variable	Focal Entity	Description
PEP Indicator	Party (customer associated with alert)	Boolean variable which signifies if the party is a Politically Exposed Person (PEP). This variable only applies to individuals or persons. This variable is a known high risk characteristic of a person.
Citizenship Country Code	Party (customer associated with alert)	Country of citizenship of the party. This variable only applies to individuals or persons.
Residence Country Code	Party (customer associated with alert)	Country of residency of the party. This variable only applies to individuals or persons.
Party AVG Monthly Transaction Amount	Party (customer associated with alert)	Monthly numeric average of total transaction amounts each month. The number of months used to compute the average depends on the analysis. For example, total monthly transaction amounts could be averaged over 3, 6, or 12 months.
Party AVG Monthly Transaction Count	Party (customer associated with alert)	Monthly numeric average of total transaction performed each month. The number of months used to compute the average depends on the analysis. For example, total monthly transaction counts could be averaged over 3, 6, or 12 months.
Party AVG Monthly Wire Amount	Party (customer associated with alert)	Monthly numeric average of total wire amounts performed each month. The number of months used to compute the average depends on the analysis. For example, total monthly wire amounts could be averaged over 3, 6, or 12 months.

Table 1. Key variables used in customer segmentation modeling example.

The variables used in this demonstration are a sample of potential variables which could be used based on the profile of the bank or financial institution doing the analysis, the availability of data, and the capabilities of the transaction monitoring system. Additional variables can be used assuming the values relate to the entity monitored by the alerts, in this case, a bank customer. Many of the variables used in this example are known high risk attributes of the entities in the analysis. True eureka moments can occur for unknown data correlations but depend entirely on the nature and content of the data. The key point to understand is that the values used as predictors in the decision tree represent potential data points which can be used to classify the entities of interest in the analysis (alert or party/customer of alert).

Purpose of Analysis

The goal of this analysis is to identify and understand risky segments of customers for productive alerts. In this analysis, false positives are not removed through data filters because the decision tree should perform the differentiation of data and the analyst can visualize productive versus nonproductive alerts. Theoretically there should be a clear path of productive versus nonproductive alerts if enough data values, and combinations of data values can be identified to each response value. The analysis hopefully shows specific segments such as the following:

1) Organizations with monthly aggregate wire transactions greater than $50,000, with high risk products, which do business in risky industries.
2) Individuals who have a risky occupation such as a real estate broker or jeweler, who are citizens from a high risk jurisdiction, that are cash intensive customers.

These are example combinations of attributes of customers, other combinations will exist and depend entire on the source data.

Building the Decision Tree

To build the decision tree, a SAS Visual Analytics Exploration is used and the sample data described in Table 1 is loaded. The Investigation Outcome is assigned to the Response role on the right of the exploration. Important customer attributes are assigned to the predictor roles on the right. This is shown in Figure 4 below. The Investigation Outcome values will be False Positive, Investigation (NOSAR), and Investigation (SAR). False positives represent nonproductive alerts while values which lead to Investigations are productive. To add additional context, values of SAR and NOSAR for alerts which lead to Investigation provide an additional layer of what is productive but also what is extra productive in that the investigation leads to filing a Suspicious Activity Report (SAR) with FinCEN.

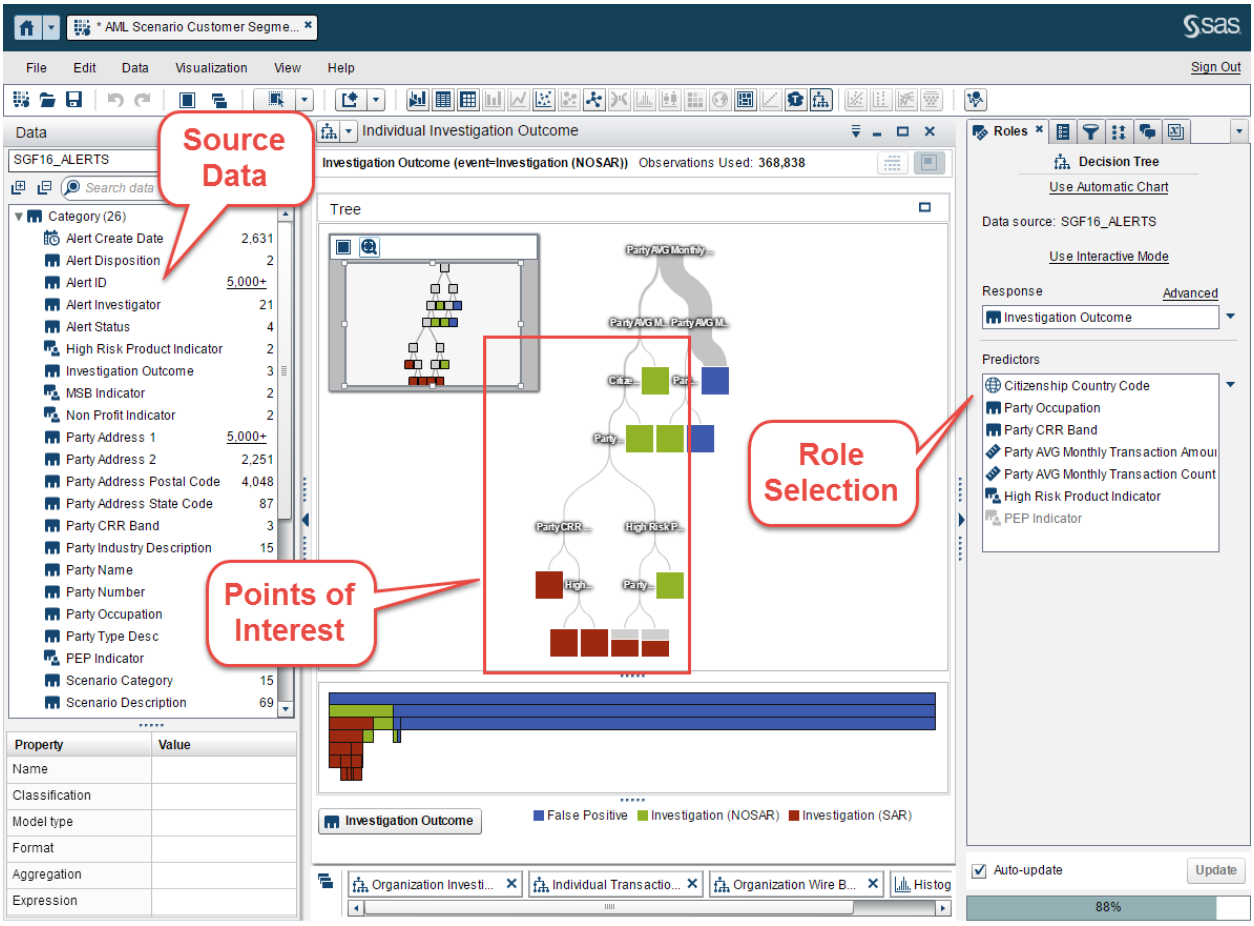

Figure 4. Overview of Decision Tree Visualization for Customer Segmentation of Productive Alerts for the INDIVIDUAL Party Type.

Once Roles have been assigned, SAS Visual Analytics produces the initial decision tree. Advanced features are enabled if your environment has SAS Visual Statistics licensed.

Two separate decision trees are built for the two very different party types, the individual person and the organization, which could be a small business or large corporation. Party type could have been used as another predictor variable but in the context of this analysis it does not make sense because both are known to be drastically different. For example, a large corporation is going to spend money much different from a person. Due to the known distinction between party types and common practice of segmenting individuals from organizations, filters were used to limit party types in two separate trees, one for individual and the other for organization. This is shown in Figure 5. Figure 4 above shows a decision tree focused on the individual. Figure 6 below shows a decision tree focused on the organization.

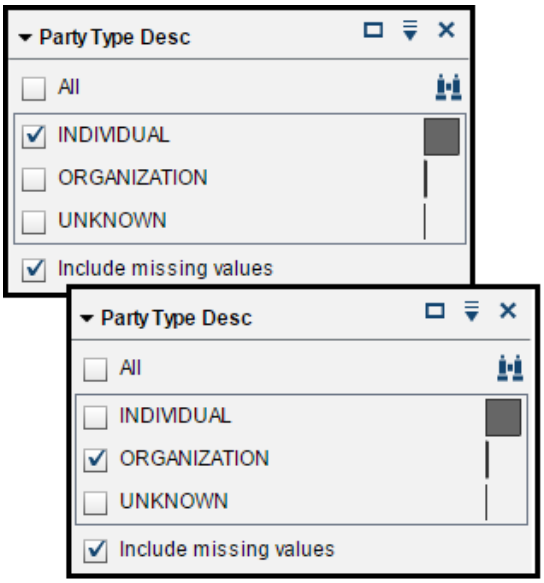

Figure 5. Filtering Party Type to Split Key Differences.

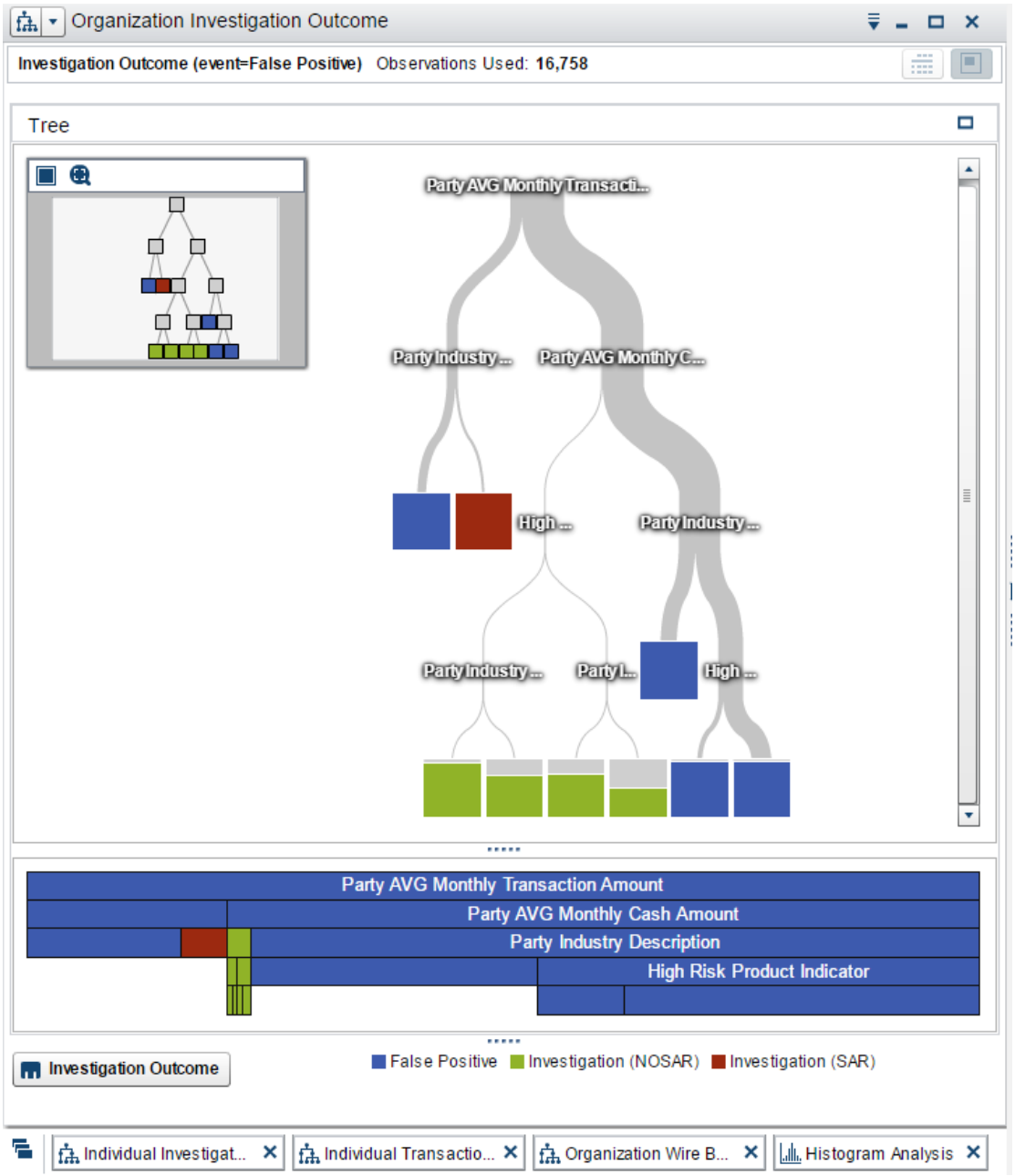

Figure 6. Decision Tree filtered to Organization party types, which represent small and large corporations.

Analyzing the Decision Tree

This analysis will focus on the INDIVIDUAL analysis shown in Figure 4 above. The following steps are taken to analyze this decision tree.

1) Starting from the top node of the decision tree, understand the distribution of response values across all data available to the decision tree.
    a. This example analysis demonstrates the common skewed nature of false positives within current AML transaction monitoring systems.

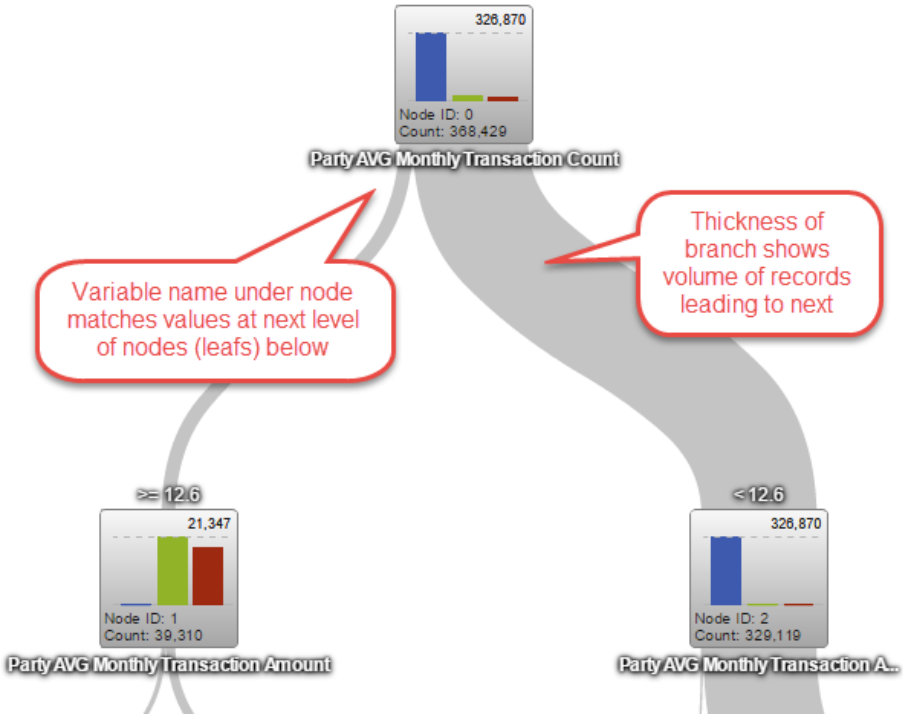

Figure 7. Top of INDIVIDUAL decision tree differentiating average monthly transaction volume.

2) The purpose of this analysis is to determine combinations of variables which indicate risky attributes of customers that lead to productive work. Therefore, this analysis should follow nodes which have a higher ratio of green and red bars, which represent a higher volume of productive alerts. Following the left branch, the analysis leads us to understand that customers with average monthly transaction volume greater than or equal to 12.6 transactions lead to a higher ratio of productive alerts. This is shown above in Figure 7.

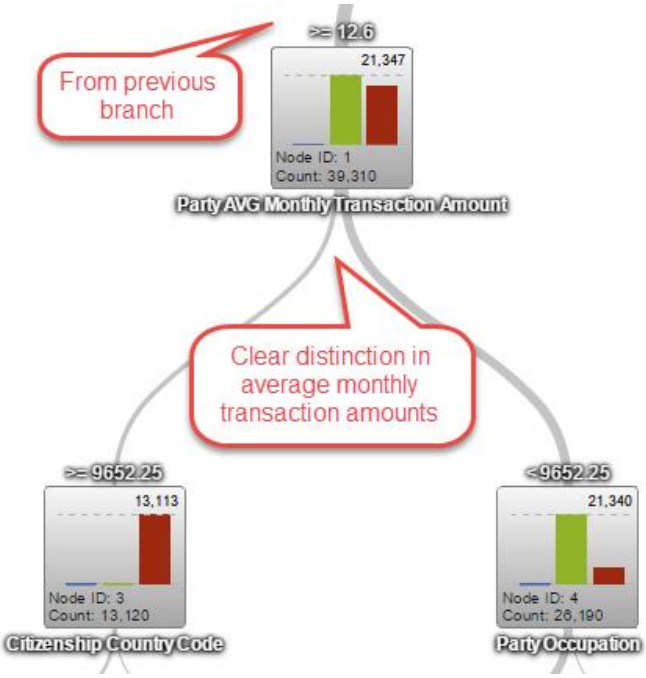

Figure 8. First productive node of INDIVIDUAL decision tree which breaks down average party transaction amount.

3) Figure 8 above continues down the decision tree. Branches shown here differentiate additional transaction behavior for individual persons. In this section of the decision tree there is a breakdown of average transaction amounts right below a key threshold of $10,000, which is a common reporting threshold for financial institutions requiring documentation of the transaction. Criminals like to avoid this threshold and will typically avoid going over this threshold, even in aggregate amounts, in order to remain inconspicuous. This section of the decision tree shows more productive alerts which result in a SAR filing for transactions greater than or equal to $9,652.25. So far the segment covers transactional spending for individual customers with an average monthly transaction count great than or equal to 12.6 and an average monthly transaction dollar amount great than or equal to $9,652.25.

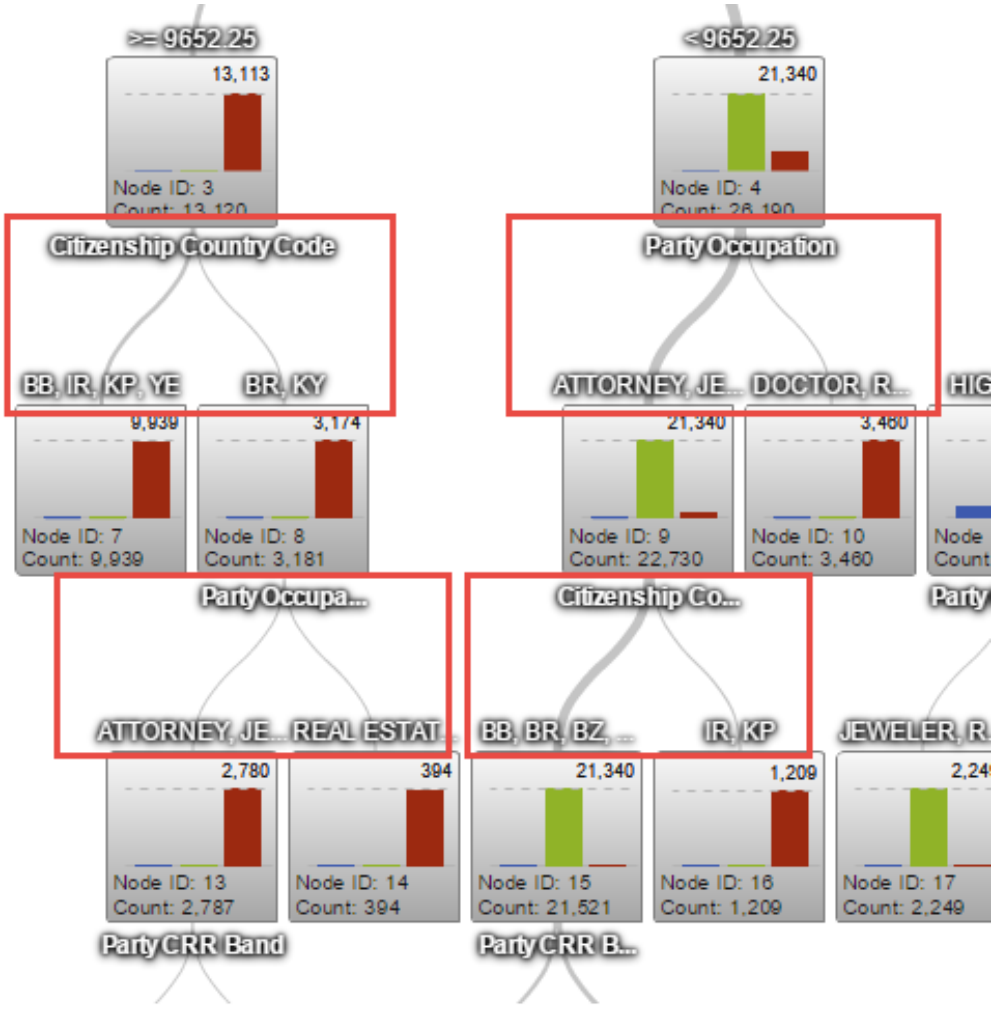

Figure 9. Further breakdown of INDIVIDUAL attributes in decision tree.

4) The rest of the decision tree breaks down into further branches of attributes which provide more complex segments. This is shown in Figure 9 above. The complexities in these branches are regulated by the decision tree options such as maximum branches, maximum levels, lead size, predictor bins, and pruning aggressiveness.

Initial Conclusions Drawn from Analysis

An example complex segment identified from this example analysis would be the following:

1) Individual customers
2) Average monthly transaction count great than or equal to 12.6

3) Average monthly transaction dollar amount great than or equal to $9,652.25
4) Citizenship from Yemen, Iran, Barbados, North Korea, Brazil, or the Cayman Islands
5) Occupations such as Attorney, Jeweler, or Real Estate Broker

To identify this complex segment, values from leafs were identified following the volume of alerts which lead to an actual SAR filing (the red bars). These are the most productive alerts. These factors could be included as a part of a complex transaction monitoring scenario, a model risk validation and tuning process, broken out into individual components for a customer risk ranking model, or any number of analytical needs which depend on segmenting the customer population. The analyst should validate findings, confirm results by simulating decision logic in the source data, then expand further as needed. For example, continuous variables such as average monthly transaction amount could be modeled even further to find the optimal threshold. This would most likely round the thresholds up to more round numbers for simplicity.

The reason for focusing on only productive alerts is because there is a confirmed reason for suspicious activity and the customers associated with these productive alerts have something which has some degree of risk associated that triggered the productive investigation. One key point is that the analysis does not remove non-productive alerts because the logic behind decision trees needs to be aware of values which lead to this response value as well. Another key point to the analysis in this paper is that the data has been simulated in order to tell a story that would be a common approach used within the banking industry. Even though data is simulated, it is representative of common risk segments. Every bank's customer profile is different, therefore this sample analysis data should be taken with a grain of salt.

UNDERSTANDING CUSTOMER TRANSACTION PROFILE

Understanding transaction behavior for customers is another common task for AML analysts. Decision trees are useful for visualizing what attributes differentiate money movement amongst customers of a bank. Rather than using a discrete response variable, this example analysis will use a continuous numerical value to visualize a histogram on each leaf of the decision tree. This allows the analyst to understand how transactions occur for each customer segment across the entire population of AML alerts. An overview of this example is shown in Figure 10 below. A key difference between this example analysis and the previous example is that the Investigation Outcome is not filtered, but rather displayed on the decision tree itself as an additional predictor. The very top of this decision tree example immediately differentiates productive versus nonproductive alerts.

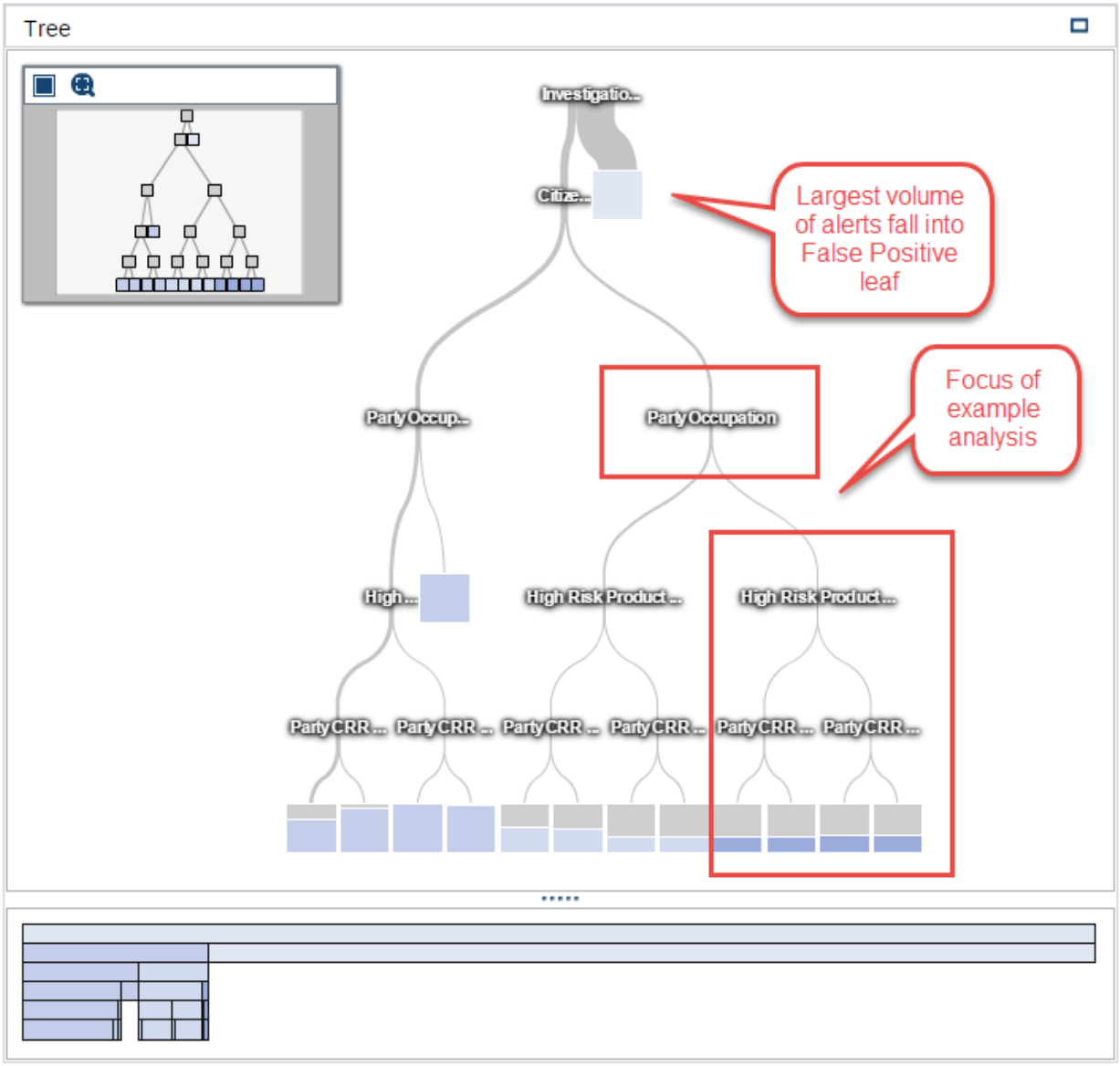

Figure 10. Overview of Individual Transaction Profile Decision Tree.

At the top of the decision tree the Investigation Outcome differentiates productive and nonproductive alerts. Productive alerts follow the left hand side of the decision tree and expand into further complex segments. The initial right branch ends with all false positives into a single leaf node. This is shown in Figure 11.

Within nonproductive alerts, the average monthly transaction amount is very skewed, indicating very low average transaction amounts for the majority of alerts in the sample data of this analysis.

Within the productive alerts, country of citizenship begins to distinguish spending habits in the sample data of this analysis. The average transaction amount shows a slightly more uniform distribution with additional average transaction amounts in larger bins in the histogram. This is shown in Figure 11 to the right. As the analysis follows this branch, the distribution of average transaction amounts will begin to breakdown into further segments.

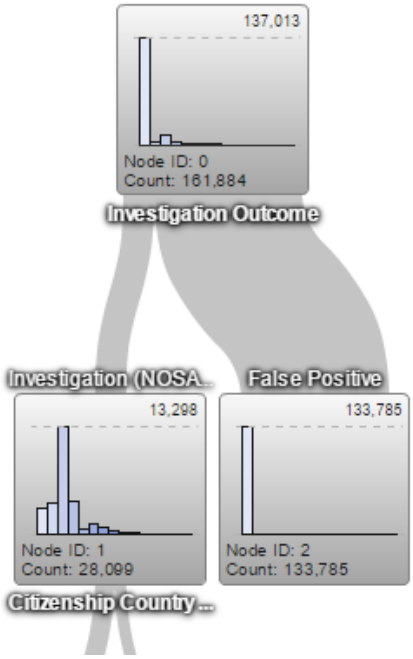

Figure 11. Average transaction amount is extremely skewed in the false positive leaf node.

After citizenship, the Party Occupation differentiates transaction behavior. After the second level branches, the decision tree expands further into additional areas of risk such as the high risk product indicator and the Party Customer Risk Ranking. The focus of this example analysis follows the highlighted sections of the decision tree shown in Figure 10 above. The following leafs represent the breakdown of segments within the highlighted section.

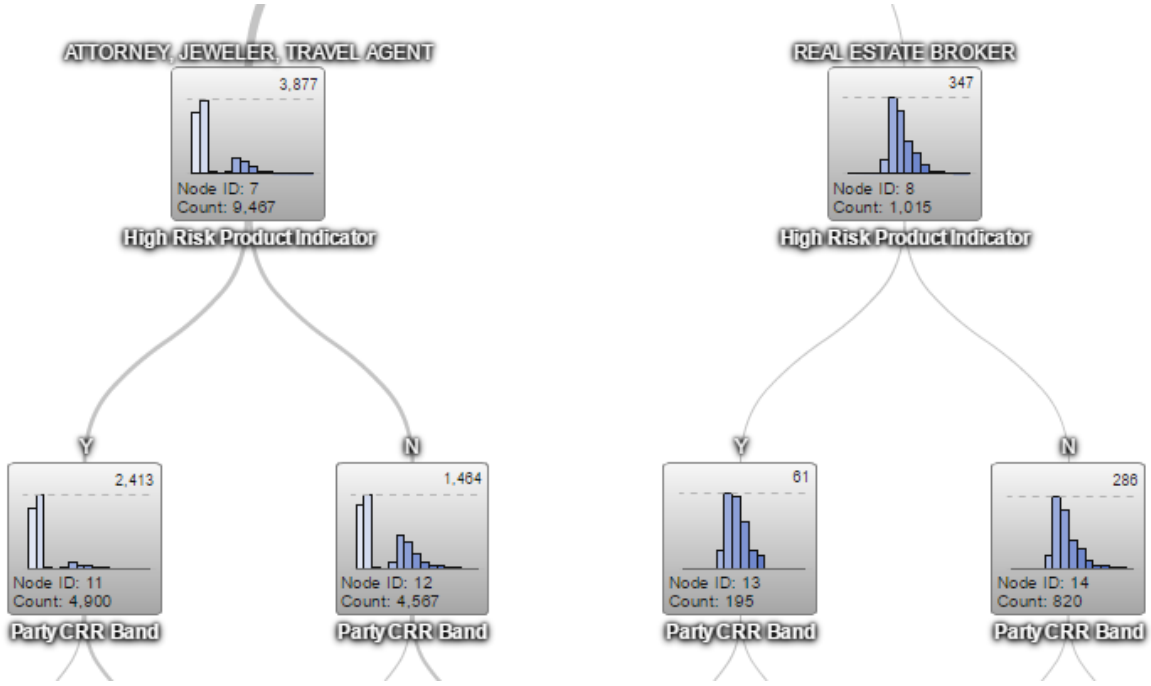

Figure 12. Segments which show average transactional behavior for productive alerts and risky customer segments.

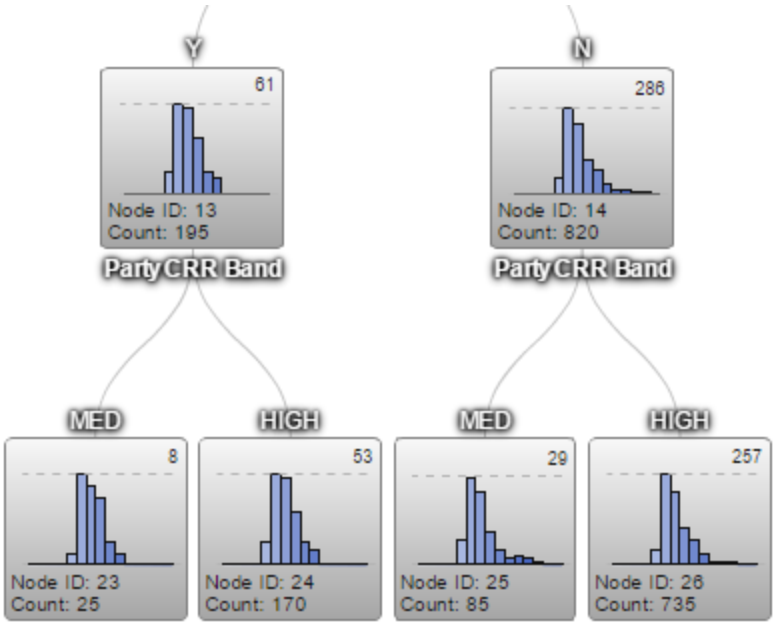

Figure 13. Lowest leaf nodes within productive alerts which follow the occupation, and high risk product indicator values.

The purpose in this analysis is to understand average transaction amounts distributed amongst productive alerts and risky customer segments. Additional analysis and growth strategy tuning should be performed in order to validate the values shown in the decision tree help predict the response values show. Creating visualizations from each node is a key functional component SAS Visual Analytics provides which makes the additional analysis efficient.

CONCLUSION

Within SAS Visual Analytics, decision trees are useful tools to help data analysts explore and understand relationships within data. Similar to other modules available in data explorations and within SAS Visual Statistics, decision trees enable and empower knowledgeable business analysts to satisfy curiosity and harness the power of SAS without having to earn a graduate degree in statistics.

Throughout this paper we have explored the many ways that decision trees can be customized and analyzed to help answer questions, identify relationships, differentiate patterns, and help highlight the signal buried within data. The important takeaway is to remember that each situation will present different opportunities, and some data will have more notable and interesting patterns, while some data will have none. Decision trees are but one possible tool for identifying those patterns, and that means that sometimes they will be useful and provide compelling results, especially when the data is willing, but sometimes they will not be the best option. A dose of pragmatism and a big picture perspective are helpful to keep all analysts mindful of the painful reality of data.

To be sure, as a dynamic and versatile tool, implemented in a customizable and accessible interface within SAS Visual Analytics, decision trees can be the first stop for many analysts seeking to better understand the patterns in the data. As much as the data has patterns, decision trees will help ensure that SAS is giving every user the power to know.

RECOMMENDED READING

- *SAS® Visual Analytics User's Guide*

CONTACT INFORMATION

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

Stephen Overton
Zencos Consulting
(919) 341-9667
soverton@zencos.com
http://stephenoverton.net
https://www.zencos.com/

Ben Murphy
Zencos Consulting
(734) 335 – 0453
bmurphy@zencos.com
https://www.zencos.com/