Paper 3622-2019

Unsupervised Learning for Anomaly Detection in Securities Data

Chris A. Robinson and Laura Rudolphi, Wells Fargo Bank, NA

ABSTRACT

An all-too-common practice in data analytics is the use of pre-defined business groupings to drive analyses and subsequent decision-making. Product categories drive assortment layouts in retail, but are they really indicative of what shoppers buy at the same time? Geographies define divisions for sales-oriented companies, but do customers in neighboring states really behave the same? In the Compliance Technology & Analytics group at Wells Fargo, account types from an existing vendor solution were driving the peer-to-peer comparisons for the anti-money laundering (AML) transaction program that covers our securities business. Accounts of the same type, or peer group, are compared to one another to identify anomalies and "out-of-bounds" behavior for AML alerting. However, not all accounts in the same peer group are created equal, resulting in sub-optimal comparisons and less valuable alerts for our investigative team. Using SAS® Enterprise Miner™, we built more meaningful account peer groups with k-means clustering. With the new clusters, we have an apples-to-apples comparison of accounts when performing transaction monitoring and anomaly detection.

INTRODUCTION

Compliance Technology and Analytics (CTA), (formerly Financial Crimes Analytics), is part of the broader Wells Fargo Corporate Risk Compliance team. The team coordinates business data intelligence from a data and analytics perspective for customers, transactions, products, channels, and businesses. Six teams make up CTA, including Financial Crimes Surveillance, Broker-Dealer Surveillance, Compliance Analytics, Compliance Solutions, CTA Ops & Financial Crimes Solutions, Business Data and Architecture.

The Broker-Dealer Surveillance team within CTA provides trade, anti-money laundering (AML), and e-comms surveillance, supporting Wells Fargo Clearing Services (retail brokerage), Wells Fargo Securities (institutional brokerage), and the Financial Institutions Group (correspondent banking). Relevant to this topic, the team conducts automated AML transaction monitoring for the WFS customer population which includes a wide assortment of industries and firm sizes. The range of dynamics in this population presents the core challenge discussed in this paper.

Wells Fargo Securities is the trade name for the capital markets and investment banking services of Wells Fargo & Company and its subsidiaries. It includes Wells Fargo Securities, LLC (member of NYSE, FINRA, NFA, and SIPC) and Wells Fargo Prime Services, LLC (member of FINRA, NFA and SIPC). WFS headquarters are in Charlotte, NC, with international offices in London, Hong Kong, Singapore, and Tokyo.

As an institutional broker dealer, WFS delivers a comprehensive set of capital markets products and services to customers, including originating and distributing public debt and equity, hedging interest rates, commodity and equity risks, advising on mergers and acquisitions, and originating structured lending facilities and municipal bonds. WFS also has more than 250 professionals providing research and economic reporting, covering more than 2,400 securities across every major sector of the economy.

WFS TRANSACTION MONITORING OVERVIEW

Currently, WFS Transaction Monitoring compares accounts at the peer group level. Bank relationship managers assign these peer groups when an account is opened based on a

company's high level attributes or business type, and examples include government entities, for profit companies, and non-profits. The AML monitoring stream for this line of business uses the peer groups to detect anomalous activity by comparing an account's behavior to that of its peers over a set time period. While the assignment of these peer groups is far from arbitrary, their subjective nature can impact anomaly detection by comparing accounts with dissimilar transactional behavior and customer size. Based on the nature of the existing categorizations, a customer would always exist in the same peer group it was initially assigned to, regardless of changes in its relationship with the bank.

Consider a single peer group as example. Two example accounts in this peer group, based on subjective assignment, include a large life insurance company and a benefit plan at a university. When considering their relationships with the bank, both customers have similar attributes, but upon a deeper dive of their transactions, it is evident that their behavior should not be compared, given the sheer volume transacted of one compared to the other, as well as the different financial instruments transacted, investment goals, and products being used.

DATA BACKGROUND

For the initial clustering pilot, we created a data set with over 500 variables that can be categorized as follows:

- Transaction Variables – We used different time periods (1/3/6/9/12 months) and compared various transaction types, including credits and debits, as well as more granular variations. In addition to counts and amounts, we analyzed other statistically derived values by customer, including the mean, maximum, and standard deviation.
- Customer Attributes – As a bank, we are required to gather certain types of information for all of our customers, including information indicative of their risk level. Other attributes include country and whether the customer is an individual or a business.
- Account Attributes – Each customer can have multiple accounts. We looked at attributes associated with each, including other non-WFS accounts at the bank.

Data preparation was completed using Base SAS®.

ANALYSIS

We used SAS® Enterprise Miner™ to perform variable selection and clustering. This paper will not discuss the technical details of the clustering; specifics have been thoughtfully outlined and curated in various other papers and presentations and should be referenced for more specific questions regarding the statistical nuances of each.

After preparing the data, we imported it as a SAS data set into Enterprise Miner and performed the following steps.

VARIABLE SELECTION

We used the default parameters for the Variable Selection node, but changed the maximum missing percentage to 10%. The default for this parameter is 50%.

Figure 1 Variable Selection Parameters

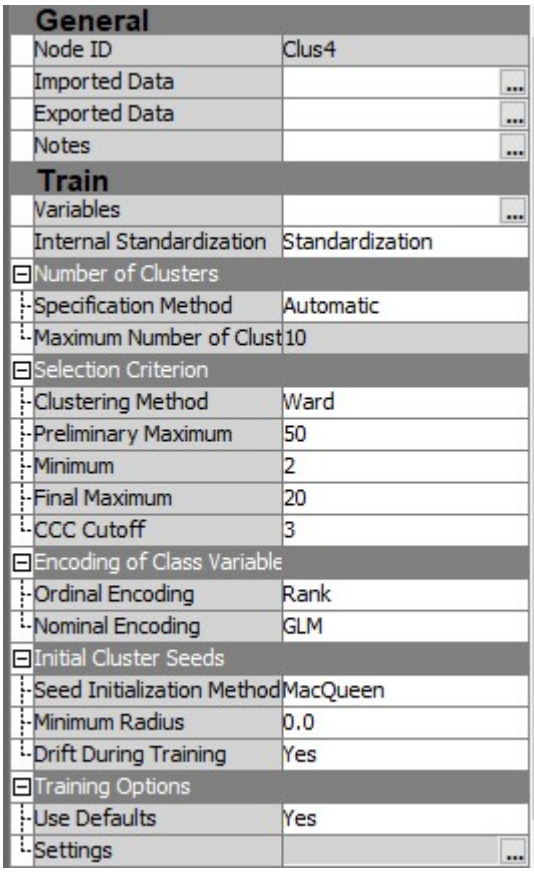
General	
Node ID	Varsel8
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Max Class Level	100
Max Missing Percentage	10
Target Model	R and Chi-square
Manual Selector	...
Rejects Unused Input	Yes
Bypass Options	
Variable	None
Role	Input
Chi-Square Options	
Number of Bins	100
Maximum Pass Number	10
Minimum Chi-Square	3.84
R-Square Options	
Maximum Variable Number	20
Minimum R-Square	0.01
Stop R-Square	5.0E-4
Use AOV16 Variables	Yes
Use Group Variables	Yes
Use Interactions	No
Use SPD Engine Library	Yes
Print Option	Default

Figure 1 Variable Selection Parameters

CLUSTERING

We used the default parameters for the Cluster node. Standardization was incredibly important in this analysis since transaction amounts and counts, as well as customer and account attributes, are all on different scales. We used the MacQueen Seed Initialization Method, which recalculates the centroid every time an iteration is completed, but also every time a data point changes its cluster. This results in many more centroid calculations than Lloyd's algorithm.

Figure 2 Clustering Parameters



General	
Node ID	Clus4
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Internal Standardization	Standardization
Number of Clusters	
Specification Method	Automatic
Maximum Number of Clust	10
Selection Criterion	
Clustering Method	Ward
Preliminary Maximum	50
Minimum	2
Final Maximum	20
CCC Cutoff	3
Encoding of Class Variable	
Ordinal Encoding	Rank
Nominal Encoding	GLM
Initial Cluster Seeds	
Seed Initialization Method	MacQueen
Minimum Radius	0.0
Drift During Training	Yes
Training Options	
Use Defaults	Yes
Settings	

Figure 2 Clustering Parameters

SEGMENT PROFILE

We used this node to profile the initial cluster output.

Figure 3 Segment Profile Results

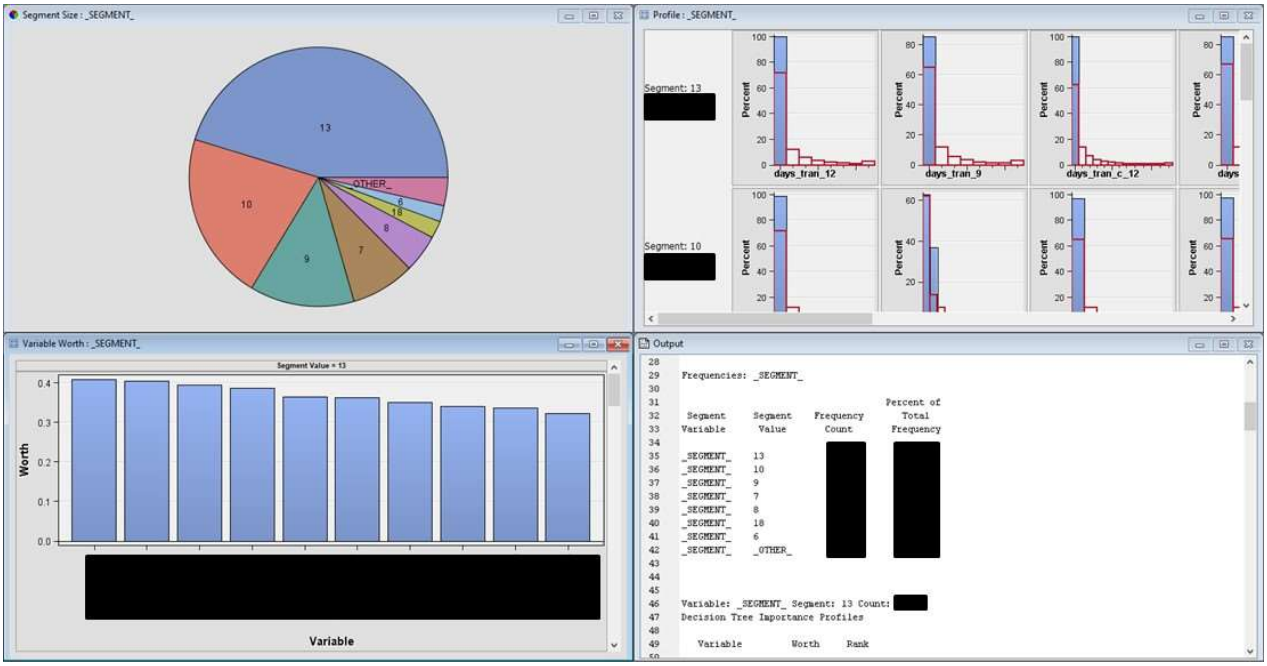


Figure 3 Segment Profile Results

SAVE DATA & SCORE

We used the Save Data node to save the output data, with cluster assignments, to a SAS data set for further analysis in Base SAS. Additionally, we used the Score node to create scoring code for Base SAS that can be more efficiently implemented in a production environment.

RESULTS

The steps and results described in this paper are the result of an iterative process. We tested numerous variable combinations, time periods, and variable selection and clustering parameters before settling on a final approach.

SAS created 20 clusters, of which we classified seven as "major" and lumped the remaining 13 into a group called "Other," infamously known in some analytic circles as a catch-all or junk cluster. While there were slight variations among important variables for each of the clusters, we discovered that the volume of transactions, specifically days transacted and transaction counts, were the most important. Also important were variables such as average daily dollar volume and maximum daily transaction.

While these results may seem commonsensical, they confirmed that the manually assigned peer groups may not be effectively applied for AML monitoring. In addition, given the emphasis on transaction volume and frequency, we tested a rank-and-decile approach for peer grouping based on key volume variables and found it to be quite similar, although not as nuanced as the clustering approach.

Also striking is our comparison of the original peer groups to the new clusters. Figure 4 below shows the original peer groups on the X axis, with the percentage mix of the clusters on the Y axis. For example, old peer group #1 is comprised 40% of cluster 13, 20% of cluster 10, and so on. The variation of clusters within the old peer groups suggests that maybe the peer groups aren't a fair way to compare after all.

Figure 4 New Cluster Mix Compared to Original Peer Group

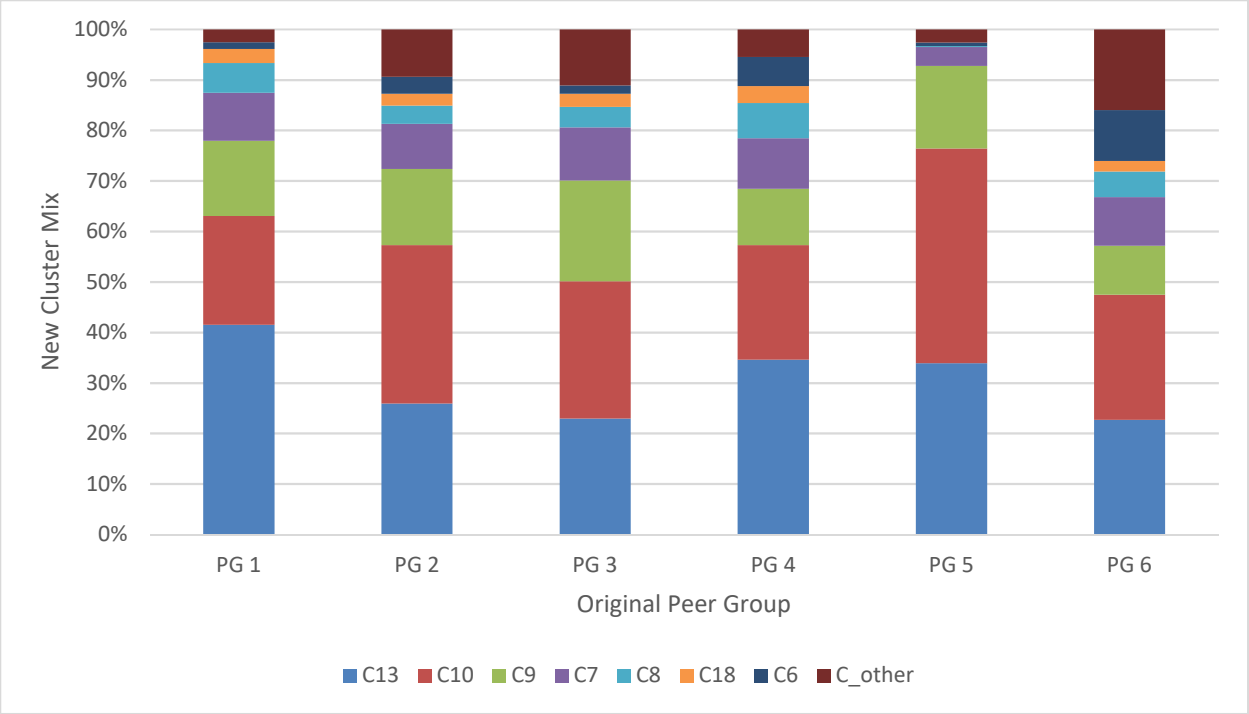


Figure 4 New Cluster Mix Compared to Original Peer Group

CONCLUSION

This clustering proof-of-concept was an incredibly valuable learning experience for the CTA team to develop a better understanding of the WFS data. At the time of this writing, steps are underway to incorporate additional data feeds for the monitoring stream that will be incorporated into future clustering iterations.

In relying on a SAS based solution to automate cluster determination, an additional benefit is the ability to re-calculate the clusters on a regular cadence that was not possible under a manual process. When a customer changes clusters, that fact can also become a valuable data point in detecting unusual behavior, which is information we did not have with static peer group assignments.

This exercise also has application to other CTA monitoring and surveillance streams and lines of business analyzed. Not all customers or accounts are created equal, and implementing program specific clustering can improve the comparisons made and enhance anomaly detection algorithms.

CONTACT INFORMATION

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

Chris A. Robinson
Wells Fargo Bank, NA
Chris.A.Robinson@wellsfargo.com

Laura Rudolphi
Wells Fargo Bank, NA
Laura.Rudolphi@wellsfargo.com