Paper SAS3377-2019

Machine Learning for Effective Surveillance with SAS® Adaptive Learning and Intelligent Agent System

Rory MacKenzie, SAS Institute Inc.

ABSTRACT

In SAS® Visual Investigator, scenario administrators can author surveillance rules to detect threats and generate highly visual alerts, which can then be investigated by analysts. Analysts then use the application to perform investigations before taking an action on the alert. Wouldn't it be great if the system could learn from the investigation that an analyst has performed in order to make the alert generation process more accurate and more up-to-date as trends change? This paper describes how SAS® Adaptive Learning and Intelligent Agent System automates the surveillance authoring process and uses machine learning techniques to adapt as new threat patterns emerge. For organizations without training data sets, we also examine how unsupervised and semi-supervised learning can be used with SAS Adaptive Learning and Intelligent Agent System to provide effective surveillance solutions.

INTRODUCTION

As data volumes increase, the task of threat detection is becoming increasingly more complex. For example, banks need to identify potentially fraudulent transactions from the billions of transactions that happen each day; police forces need to identify suspects from millions of people; and insurance companies need to identify fraudulent claims from the millions of claims. Humans can no longer manually surveil the data that is available. They are reliant on systems to narrow down the data to help them find the 'needle in the haystack'. These systems tend to be one of the following:

- Expert-driven: A domain expert defines rules or predictive models, based on their industry experience, to detect the threats.
- Data-driven: Machine learning techniques are used to identify suspicious patterns in the data (Pozzolo 2015).

While these approaches prove to be effective, they either require domain level expertise to define adequate expert-driven rules or data science skills to define the data-driven rules. Both also require regular updates as threat patterns change and evolve.

In SAS Visual Investigator, users can utilize SAS® Scenario Administrator features to define expert-driven detection strategies. This paper introduces a new feature of SAS Visual Investigator called SAS® Adaptive Learning and Intelligent Agent System. It allows users to automatically generate detection rules for their industry based on their data. As analysts use the system, it tunes and evolves the rules over time using both supervised and unsupervised machine learning techniques.

DETECTION: THE CHALLENGES

This section highlights the key challenges that users of SAS Visual Investigator typically face in their industries.

HUGE DATA VOLUMES

A bank will store billions or trillions of transactions and will process tens to hundreds of millions of new transactions each day. An insurance company will have millions of policies and will receive thousands of new claims each day that need to be processed. Police forces might have millions of crime or intelligence reports, and access to data sources such as those generated by automatic number plate recognition (ANPR). These generate new data points every second. Moreover, the ever-growing data sets typically come from multiple disparate data sources. Detecting threats in such a mass of data is difficult to do in a timely fashion.

EVOLVING THREATS

Fraudsters identify areas where they can commit fraud without detection and thereafter tend to expose that weakness until it is detected. Fraudsters are always looking for new ways to commit fraud. This evolving threat poses problems for the expert-driven detection methodology since new rules need to be written as new fraud patterns emerge. This is something that takes time and expertise to implement. There needs to be an awareness in the organization that the fraud has even taken place.

LIMITED HUMAN RESOURCES

Once a threat is identified, human resources are required to investigate it further to discern whether it is indeed a threat and what further action to take. Organizations have limited analysts available to triage generated alerts. Prioritizing these alerts by threat and ensuring low levels of false positives is a key consideration in maximizing the analyst’s efficiency. Organizations need to balance the cost of undetected fraud against the cost of processing false alerts.

Moreover, there is also a shortage of analytics talent available. Despite recent efforts to educate the workforce in data science, the demand for those human resources is higher than the supply. This has led to wages for data scientists rising by 16% per year in the USA —8 times more than average—which again prohibits the ability for organizations to hire or fund these resources (McKinsey 2016).

UNBALANCED DATA SETS

An unbalanced data set is one that is skewed toward one type of classification; there are more genuine transactions than fraudulent ones, there are more genuine insurance claims than there are fraudulent ones, and there are more law-abiding citizens than there are criminals. This is a problem for most machine learning techniques as they are not designed with this in mind, and thus sampling techniques need to be applied to optimize for rare events (Karagod 2018).

DETECTION AND INVESTIGATION WITHIN SAS VISUAL INVESTIGATOR

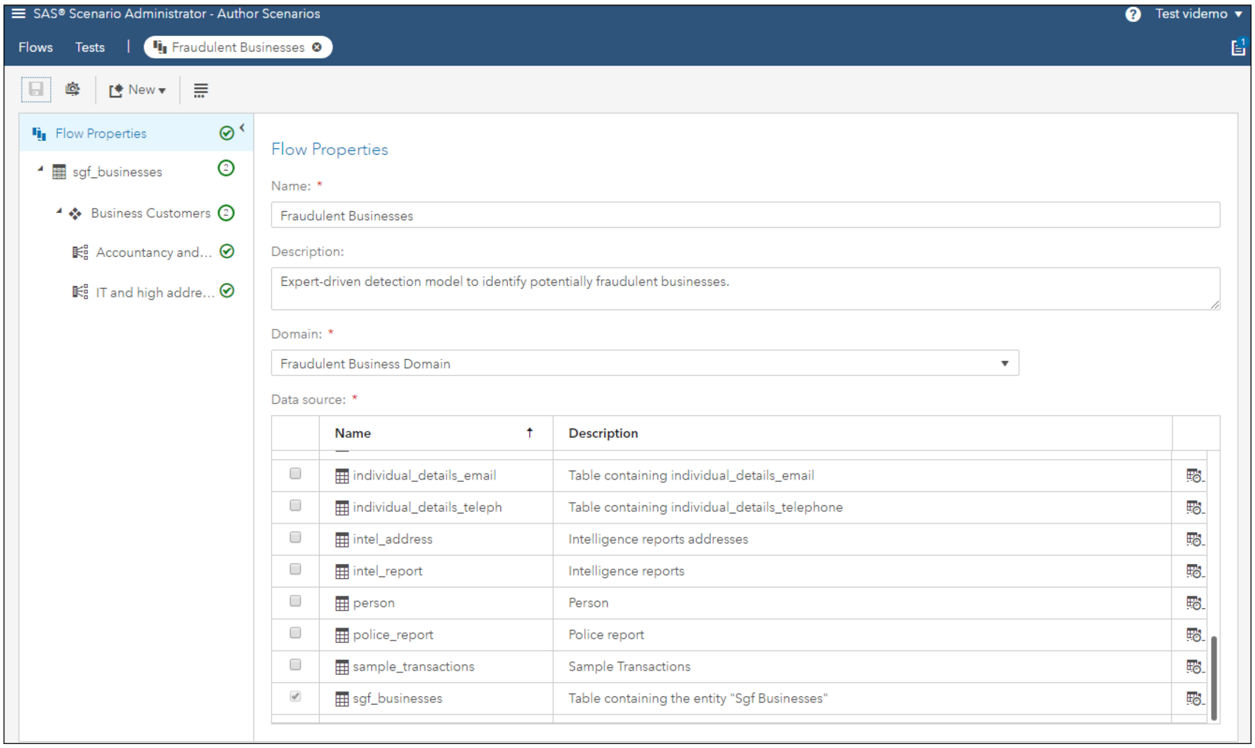
The following examples compare the expert- and data-driven surveillance approaches using the Scenario Administrator and SAS Adaptive Learning and Intelligent Agent System features of SAS Visual Investigator, with the goal of identifying fraudulent businesses. For simplicity, a synthetic fraud data set is used. It contains business accounts, business customers, and resolved entities (entities automatically derived from other data in the system) such as person, telephone, addresses, and email.

In both examples we perform detection on one table, called sgf_businesses, which contains details of the businesses. This includes name, incorporation date, country of registration, customer type, primary industry, count of previous anti-money laundering (AML) alerts, count of previous sanctions, and a fraud flag. This fraud flag is the target variable determining that the business has been previously labeled as fraud. It is for machine learning purposes and is ignored when writing the expert-driven rules.

SAS SCENARIO ADMINISTRATOR FOR EXPERT-DRIVEN DETECTION

SAS Scenario Administrator enables users to carry out expert-driven detection via a point and click interface for authoring surveillance rules and scenarios. These rules and scenarios identify threats that can generate highly visual alerts that can be triaged by analysts.

Figure 1 shows how a user would create a new flow in SAS Scenario Administrator by selecting the input table on which to build their scenarios (in our case sgf_businesses).



Flow Properties

Name: *
Fraudulent Businesses

Description:
Expert-driven detection model to identify potentially fraudulent businesses.

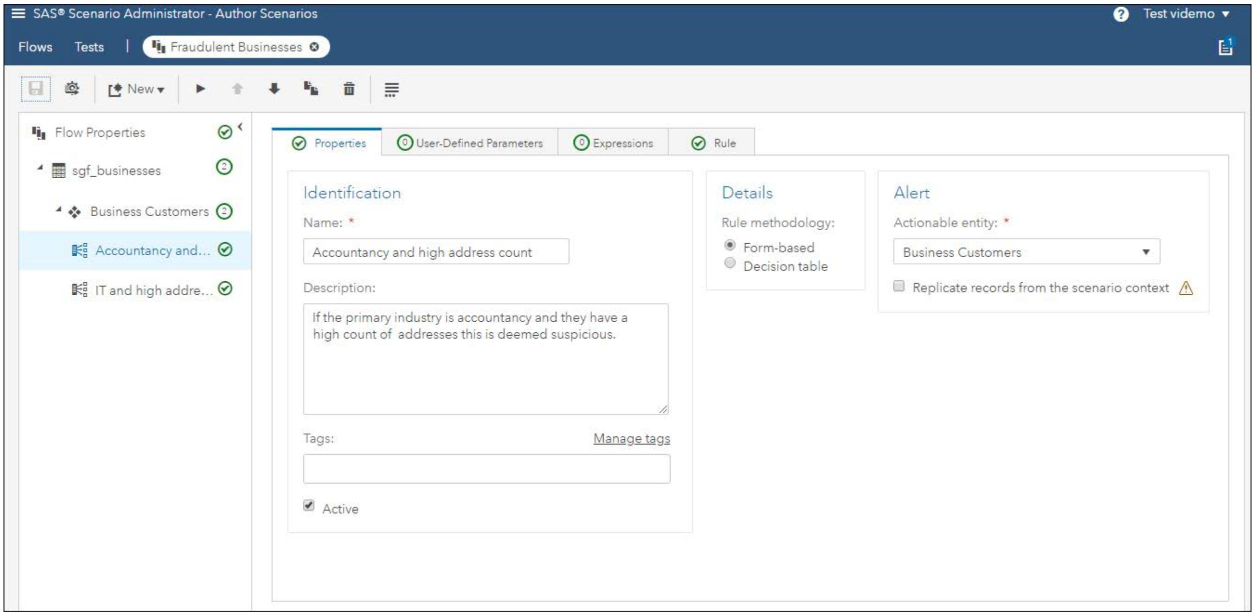
Domain: *
Fraudulent Business Domain

Data source: *

	Name	Description
☐	individual_details_email	Table containing individual_details_email
☐	individual_details_teleph	Table containing individual_details_telephone
☐	intel_address	Intelligence reports addresses
☐	intel_report	Intelligence reports
☐	person	Person
☐	police_report	Police report
☐	sample_transactions	Sample Transactions
☑	sgf_businesses	Table containing the entity "Sgf Businesses"

Figure 1. Create the Flow Using the sgf_businesses Table

When creating a flow, the user is prompted to choose the actionable entity. This is the entity that will be associated with the alert created in SAS Visual Investigator. As Figure 2 shows, in this example we want to generate the alert on the associated business customer.



Name: * Accountancy and high address count

Description: If the primary industry is accountancy and they have a high count of addresses this is deemed suspicious.

Tags: Manage tags

☑ Active

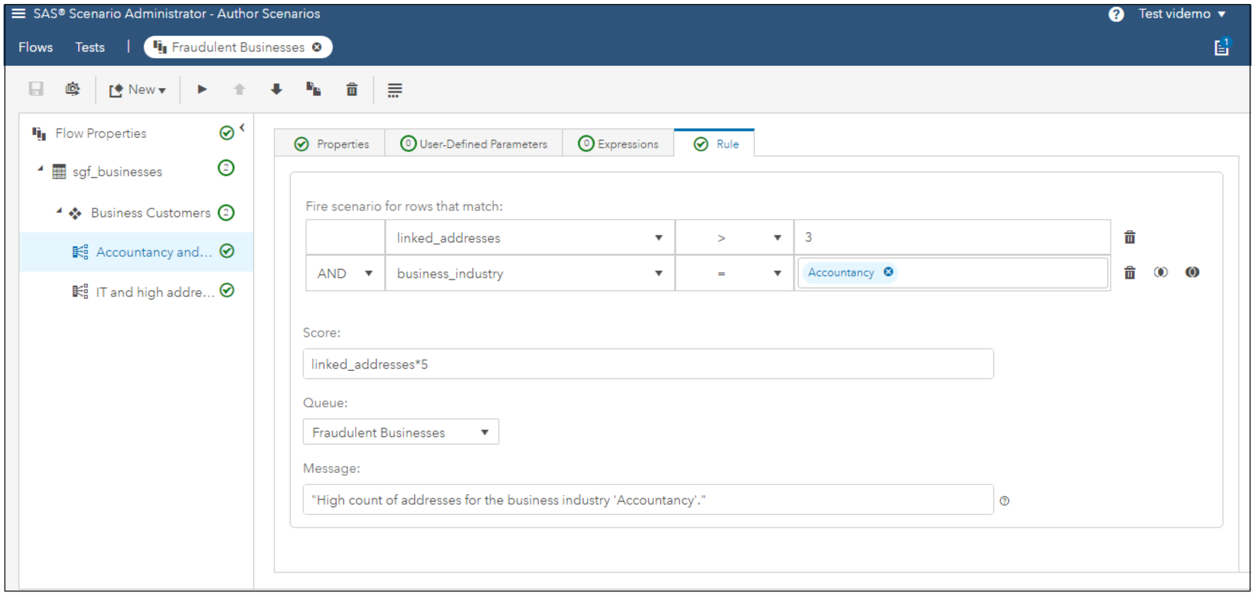
Rule methodology: ● Form-based ○ Decision table

Actionable entity: * Business Customers

□ Replicate records from the scenario context

Figure 2. Select the Business Customers Actionable Entity

For demonstration purposes, assume we know from experience that businesses with high levels of linked addresses and whose primary industry is accounting have a high chance of being fraudulent. In addition, if a business has a primary industry of IT and they have both high numbers of linked addresses and telephones they tend to be fraudulent. With this domain expertise, we can write specific scenarios using the tool and apply a score as shown in Figure 3 and Figure 4.



Fire scenario for rows that match:

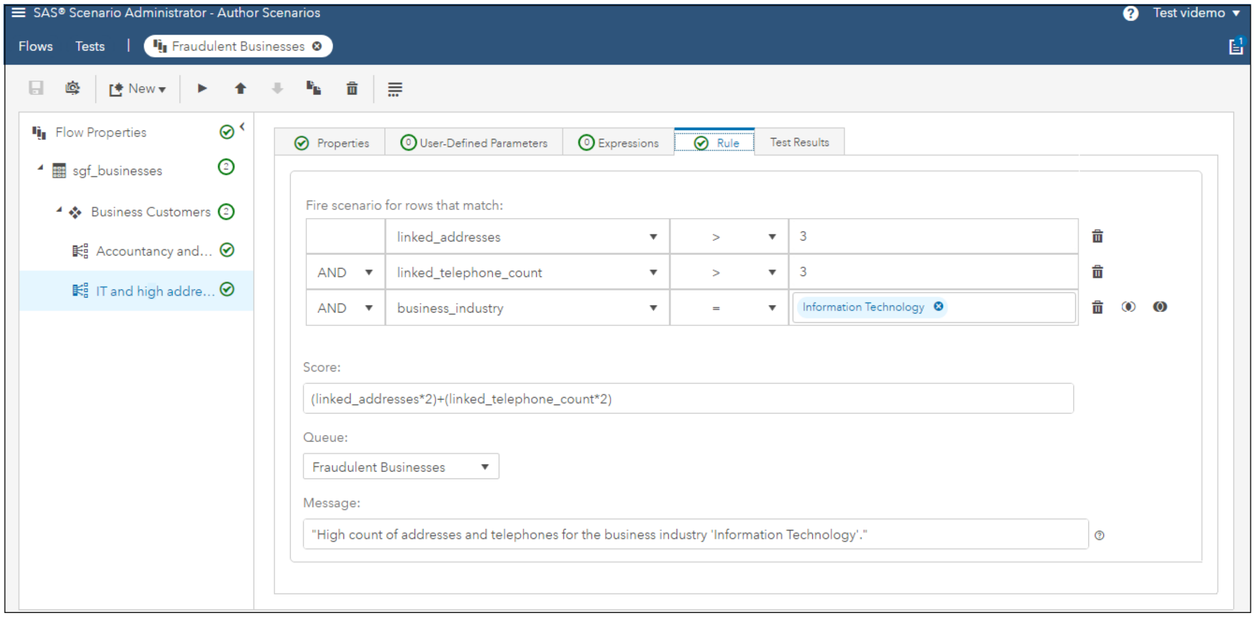
	linked_addresses	>	3
AND	business_industry	=	Accountancy

Score: linked_addresses*5

Queue: Fraudulent Businesses

Message: "High count of addresses for the business industry 'Accountancy'."

Figure 3. Create Rules to Detect Businesses with a Large Number of Linked Addresses and a Primary Industry of Accountancy



Fire scenario for rows that match:

	linked_addresses	>	3
AND	linked_telephone_count	>	3
AND	business_industry	=	Information Technology

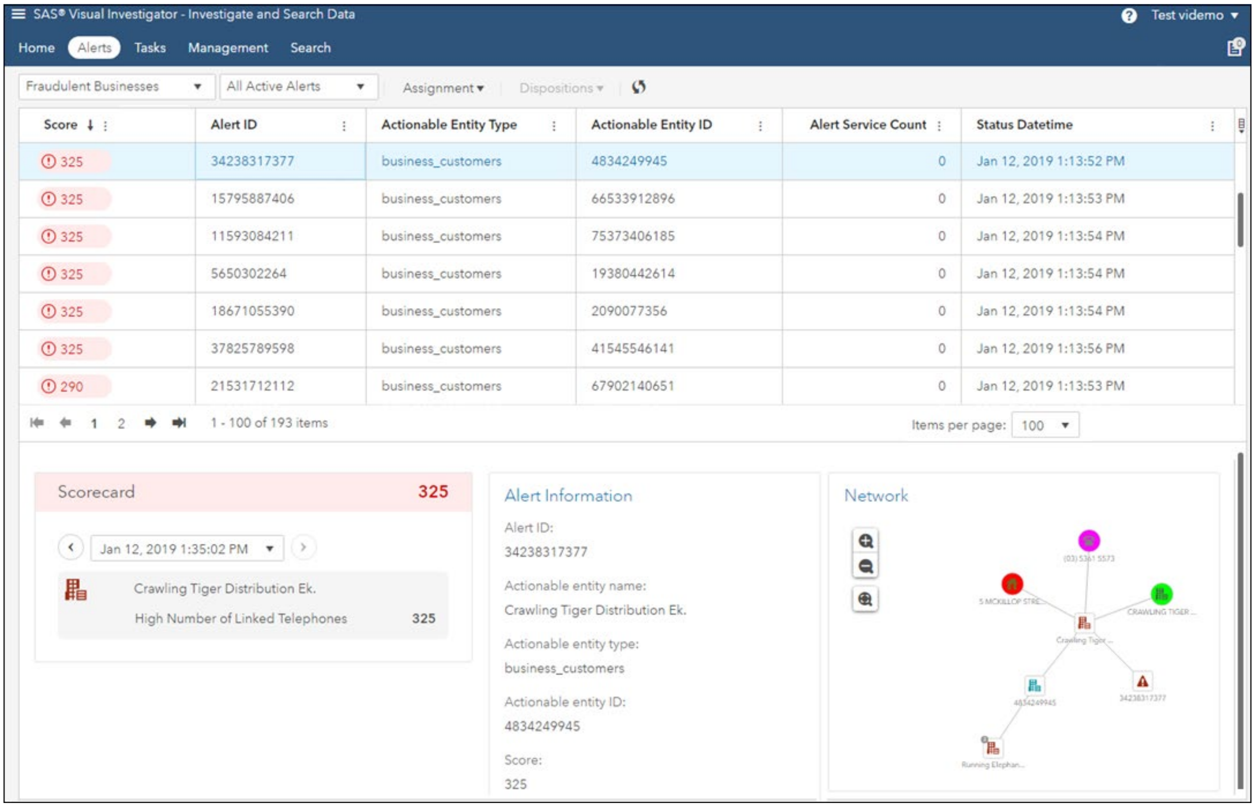
Score:
(linked_addresses*2)+(linked_telephone_count*2)

Queue:
Fraudulent Businesses

Message:
"High count of addresses and telephones for the business industry 'Information Technology'."

Figure 4. Create Rules to Detect Businesses with a Large Number of Linked Addresses and Telephones and a Primary Industry of IT

These rules can be complex, for example, you might want to write a DATA step scenario. You can add more scenarios to your flow as you uncover more approaches to fraud. When complete, the flow is published to production and used to generate alerts based on the data visible in SAS Visual Investigator. Our example of this is shown in Figure 5.



Score	Alert ID	Actionable Entity Type	Actionable Entity ID	Alert Service Count	Status Datetime
325	34238317377	business_customers	4834249945	0	Jan 12, 2019 1:13:52 PM
325	15795887406	business_customers	66533912896	0	Jan 12, 2019 1:13:53 PM
325	11593084211	business_customers	75373406185	0	Jan 12, 2019 1:13:54 PM
325	5650302264	business_customers	19380442614	0	Jan 12, 2019 1:13:54 PM
325	18671055390	business_customers	2090077356	0	Jan 12, 2019 1:13:54 PM
325	37825789598	business_customers	41545546141	0	Jan 12, 2019 1:13:56 PM
290	21531712112	business_customers	67902140651	0	Jan 12, 2019 1:13:53 PM

1 2 1 - 100 of 193 items Items per page: 100

Scorecard 325

Jan 12, 2019 1:35:02 PM

Crawling Tiger Distribution Ek.
High Number of Linked Telephones 325

Alert Information

Alert ID:
34238317377

Actionable entity name:
Crawling Tiger Distribution Ek.

Actionable entity type:
business_customers

Actionable entity ID:
4834249945

Score:
325

Figure 5. Example Alert in SAS Visual Investigator

As part of the alert generation process, associated information (such as the network around the alerted object) is brought in to enrich the alert for use as a starting point in the investigation. An analyst can use the other search and discovery tools (such as the map and timeline visualizations) to aid them in their investigation. Once complete, the analyst can apply the appropriate disposition to the alert.

SAS ADAPTIVE LEARNING AND INTELLIGENT AGENT SYSTEM FOR DATA-DRIVEN DETECTION

For organizations where the domain expertise, data science skills, or resources are not available, SAS Adaptive Learning and Intelligent Agent System provides the ability to automate building models for detection. This application looks at an organization's data and uses both supervised and unsupervised machine learning techniques to build an accurate model that is specially tuned to detect rare events such as fraud. This model can then be deployed to generate alerts in SAS Visual Investigator on existing data, as well as operate on new data as it enters the system.

These alerts are then worked by analysts, using investigative tools such as the network diagram, timelines, and maps to determine if the threat is real. The alert is then dispositioned—it's either a good alert (productive) or a bad alert (unproductive). Crucially, the intelligent agent listens to the outcome of the alert investigation and updates its learning data set with the outcome of the analysts work. After a significant number of new updates have occurred, the intelligent agent automatically retrains against the updated data set and produces a new model which, if deemed better than the last, can be published to production.

This adaptive learning process is summarized in Figure 6.

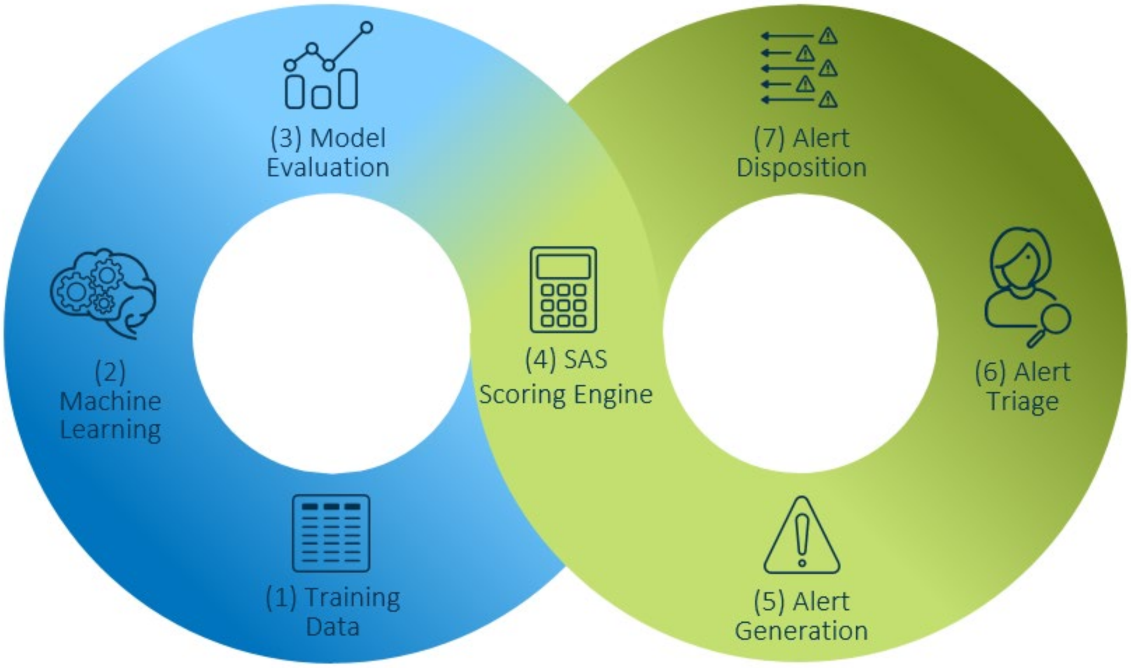
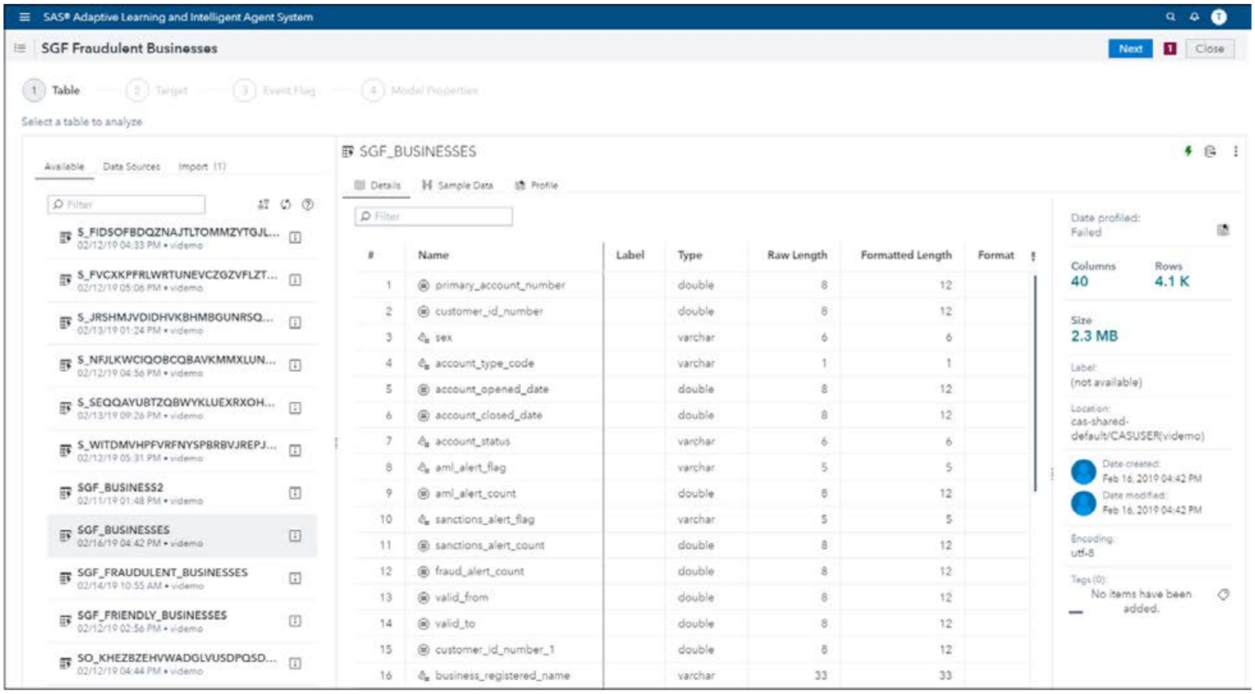


Figure 6. The Adaptive Learning Process

At stage 1 of the process we choose our training table. For demonstration purposes we select the sgf_businesses table as shown in Figure 7.



Select a table to analyze

Table	Modified
S_FIDSOFBDQZNAJTLTOMMZYTGJL...	02/12/19 04:33 PM • videmo
S_FVCXKPFRLWRTUNEVCZGZVFLZT...	02/12/19 05:06 PM • videmo
S_JRSHMJVDIDHVKBHMBGUNRSQ...	02/13/19 01:24 PM • videmo
S_NFJLKWCIQOBCQBAVKMMXLUN...	02/12/19 04:56 PM • videmo
S_SEQQAYUBTZQBWYKLUEXRXOH...	02/13/19 09:26 PM • videmo
S_WITDMVHPFVRFNYSPBRBVJREPJ...	02/12/19 05:31 PM • videmo
SGF_BUSINESS2	02/11/19 01:48 PM • videmo
SGF_BUSINESSES	02/16/19 04:42 PM • videmo
SGF_FRAUDULENT_BUSINESSES	02/14/19 10:55 AM • videmo
SGF_FRIENDLY_BUSINESSES	02/12/19 02:56 PM • videmo
SO_KHEZBZEHVWADGLVUSDPQSD...	02/12/19 04:44 PM • videmo

SGF_BUSINESSES

#	Name	Label	Type	Raw Length	Formatted Length	Format
1	primary_account_number		double	8	12	
2	customer_id_number		double	8	12	
3	sex		varchar	6	6	
4	account_type_code		varchar	1	1	
5	account_opened_date		double	8	12	
6	account_closed_date		double	8	12	
7	account_status		varchar	6	6	
8	aml_alert_flag		varchar	5	5	
9	aml_alert_count		double	8	12	
10	sanctions_alert_flag		varchar	5	5	
11	sanctions_alert_count		double	8	12	
12	fraud_alert_count		double	8	12	
13	valid_from		double	8	12	
14	valid_to		double	8	12	
15	customer_id_number_1		double	8	12	
16	business_registered_name		varchar	33	33	

Date profiled: Failed

Columns 40 Rows 4.1 K

Size 2.3 MB

Label: (not available)

Location: cas-shared-default/CASUSER(videmo)

Date created: Feb 16, 2019 04:42 PM

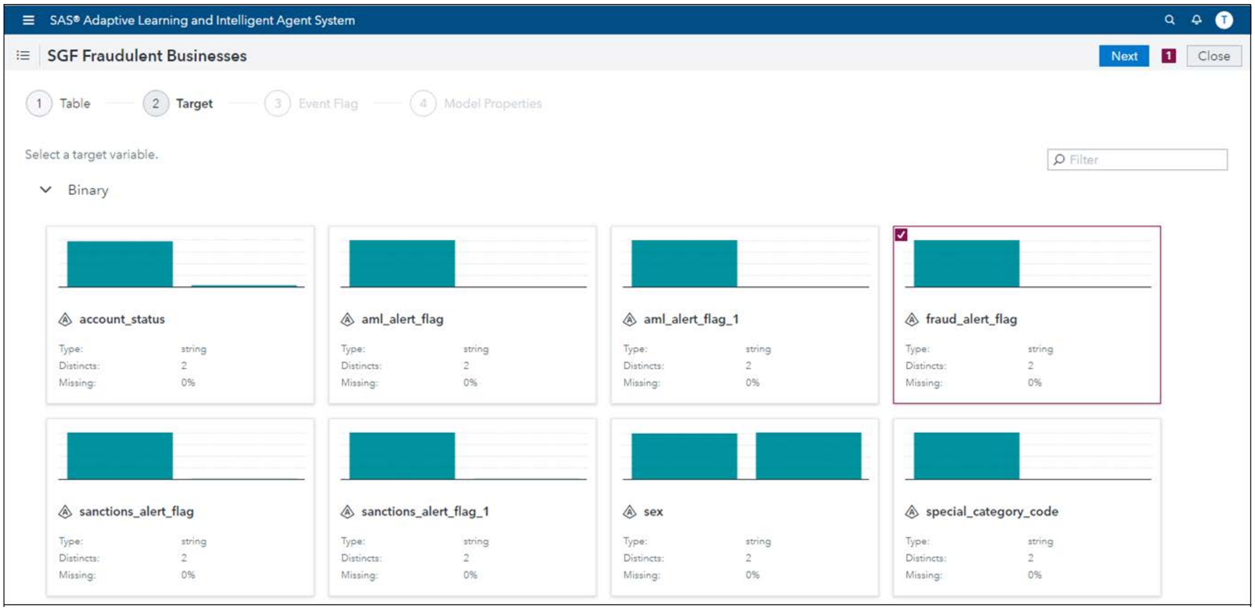
Date modified: Feb 16, 2019 04:42 PM

Encoding: utf-8

Tags (0): No items have been added.

Figure 7. Select the Training Data

After the table is selected, we need to identify the target—this is the target variable that the machine learning algorithms will use to learn from and that we are ultimately trying to predict. The intelligent agent then profiles your data and shows the binary and categorical variables as possible candidate targets. In this case we select the fraud_alert_flag variable as shown in Figure 8.


SAS® Adaptive Learning and Intelligent Agent System
SGF Fraudulent Businesses
Next
1
Close
1 Table
2 Target
3 Event Flag
4 Model Properties
Filter

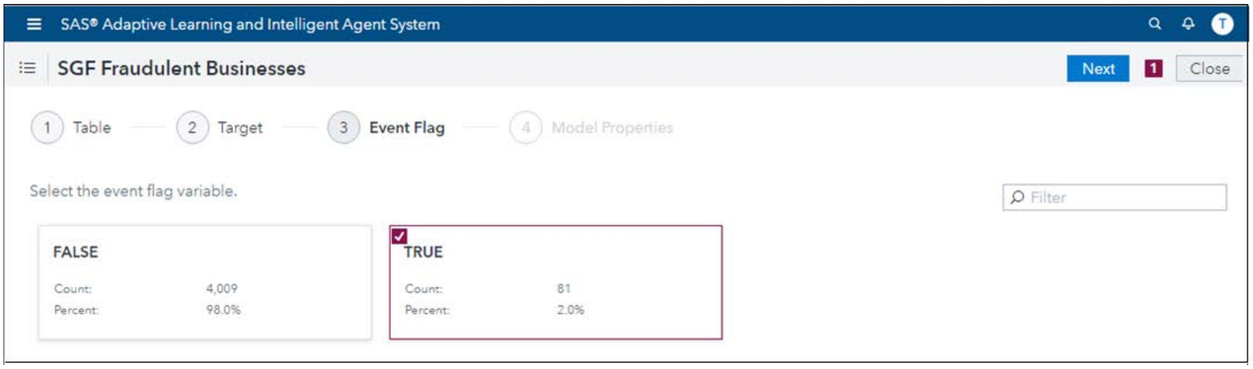
Select a target variable.

Binary

Variable	Type:	Distincts:	Missing:
account_status	string	2	0%
aml_alert_flag	string	2	0%
aml_alert_flag_1	string	2	0%
☑ fraud_alert_flag	string	2	0%
sanctions_alert_flag	string	2	0%
sanctions_alert_flag_1	string	2	0%
sex	string	2	0%
special_category_code	string	2	0%

Figure 8. Select the fraud_alert_flag Variable

Next, we select TRUE as the event flag. Note that TRUE is a rare event in this data set, representing 2% of the data. This is shown in Figure 9. This is a case of the previously discussed unbalanced data set.


SAS® Adaptive Learning and Intelligent Agent System
SGF Fraudulent Businesses
Next
1
Close
1 Table
2 Target
3 Event Flag
4 Model Properties
Filter

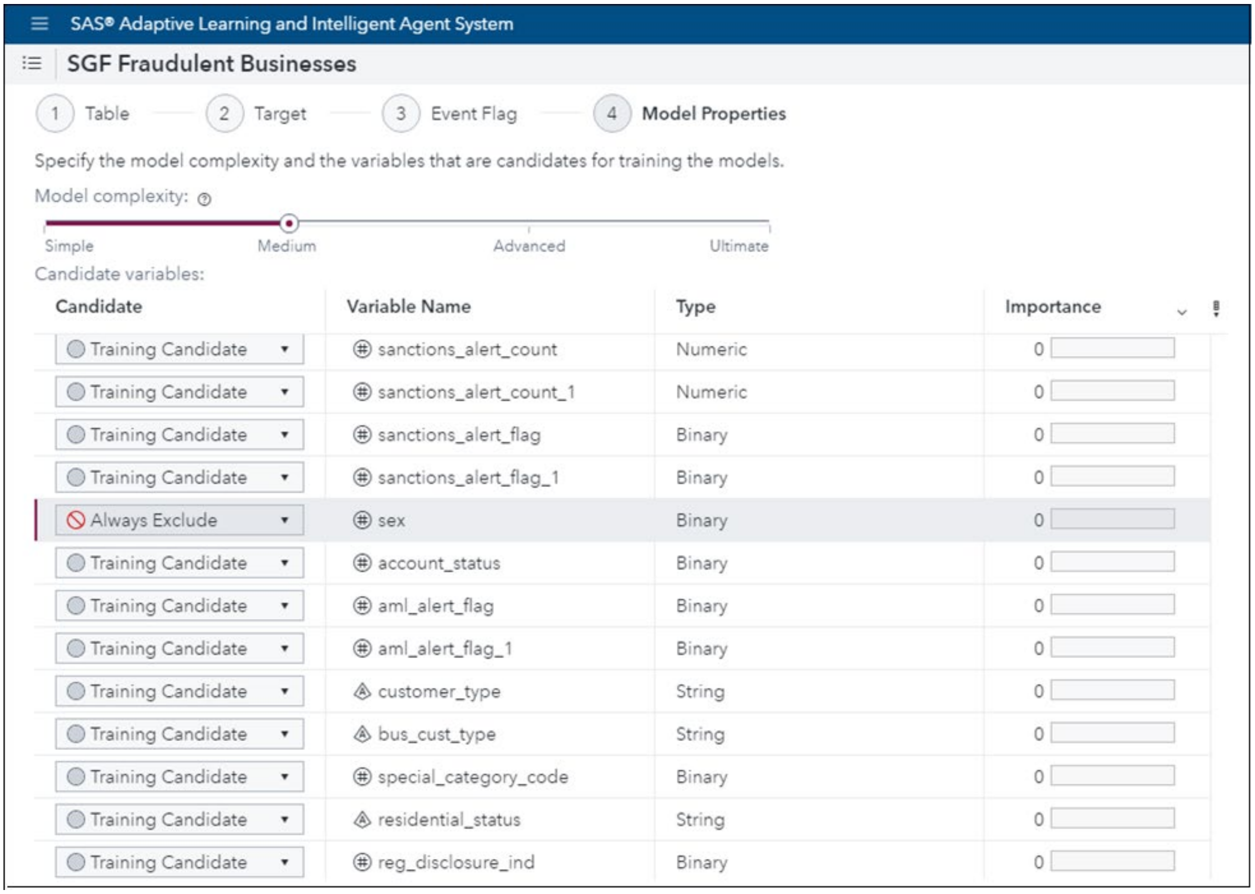
Select the event flag variable.

FALSE
Count: 4,009
Percent: 98.0%

☑ TRUE
Count: 81
Percent: 2.0%

Figure 9. Select the Event Flag

The final step before training the model is to tweak any model properties. We can assign the model complexity, which dictates the space of tuning parameters that seed the machine learning models. We can also decide which candidate variables should be included as part of the model. We can choose to always include a variable or always exclude a variable. For example, you might always exclude a variable if you want to remove race or sex as a candidate variable for compliance reasons. Alternatively, you can force the inclusion of a variable that experts think is important. This is shown in Figure 10.


SAS® Adaptive Learning and Intelligent Agent System
SGF Fraudulent Businesses
1 Table
2 Target
3 Event Flag
4 Model Properties
Simple
Medium
Advanced
Ultimate

Specify the model complexity and the variables that are candidates for training the models.

Model complexity: Medium

Candidate variables:

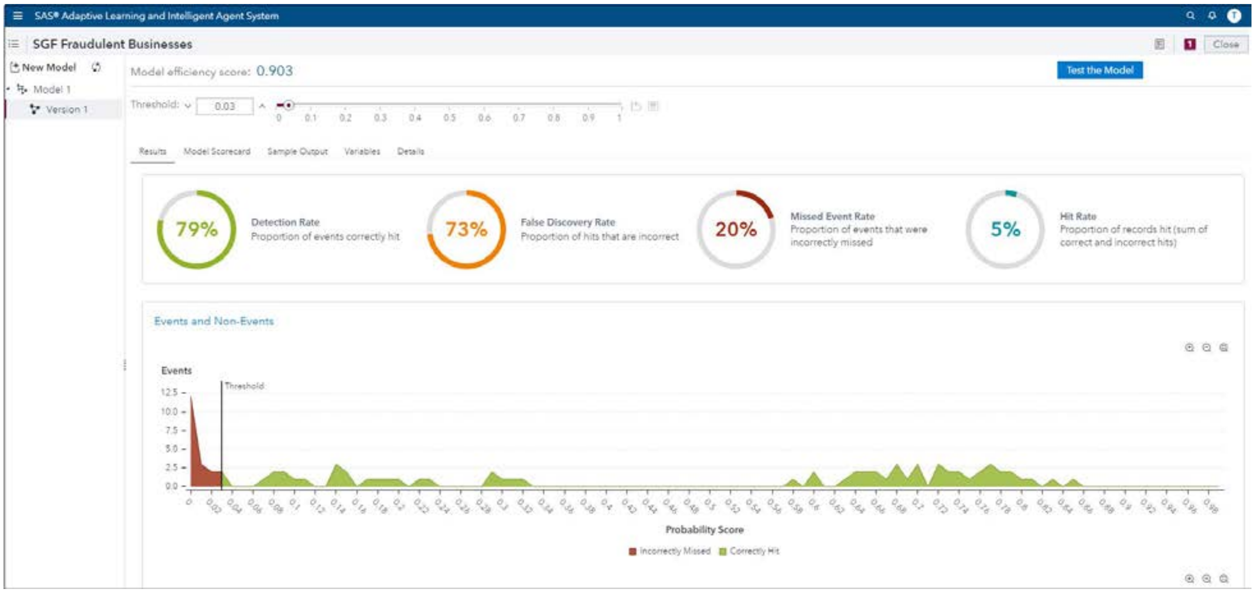
Candidate	Variable Name	Type	Importance
Training Candidate	sanctions_alert_count	Numeric	0
Training Candidate	sanctions_alert_count_1	Numeric	0
Training Candidate	sanctions_alert_flag	Binary	0
Training Candidate	sanctions_alert_flag_1	Binary	0
Always Exclude	sex	Binary	0
Training Candidate	account_status	Binary	0
Training Candidate	aml_alert_flag	Binary	0
Training Candidate	aml_alert_flag_1	Binary	0
Training Candidate	customer_type	String	0
Training Candidate	bus_cust_type	String	0
Training Candidate	special_category_code	Binary	0
Training Candidate	residential_status	String	0
Training Candidate	reg_disclosure_ind	Binary	0

Figure 10. Select the Model Complexity and Candidate Training Variable

Unless explicitly included or excluded, the intelligent agent uses an algorithm to identify the best set of variables to use in building the model. We are aided in candidate selection by a preliminary estimate of variable importance, indicating which variables are likely to be higher indicators of fraud. Note that the set of variables used might change each time the model is retrained due to the data changing.

At stage 2, once training is initiated, the intelligent agent automatically tunes hyper-parameters for each machine learning algorithm, for example, the number of trees. Once tuning is complete, the intelligent agent generates an ensemble model utilizing multiple machine learning techniques. For each model in the ensemble, the intelligent agent performs sub-sampling and partitioning to account for the natural imbalance that occurs in data sets such as fraud. The end user requires no understanding of what the intelligent agent is doing—it is a black box system.

At stage 3, we evaluate the model that the intelligent agent has produced. This is shown in Figure 11.



SGF Fraudulent Businesses

Model efficiency score: 0.903

Threshold: 0.03

79% Detection Rate — Proportion of events correctly hit

73% False Discovery Rate — Proportion of hits that are incorrect

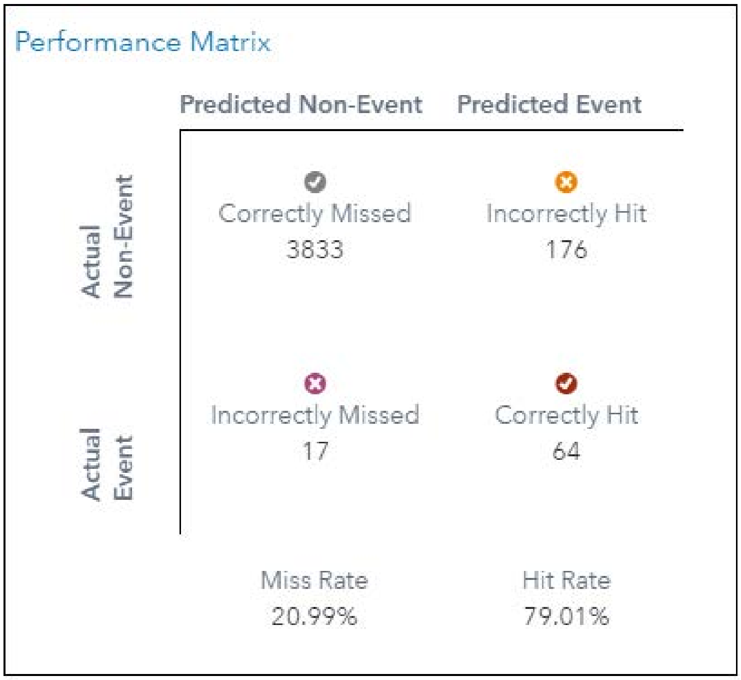
20% Missed Event Rate — Proportion of events that were incorrectly missed

5% Hit Rate — Proportion of records hit (sum of correct and incorrect hits)

Figure 11. Model Evaluation

The intelligent agent automatically calculates the optimal threshold score above which you would alert upon to maximize the detection rate, while minimizing incorrect hits or false positives. We can increase or decrease this threshold to yield higher or lower detection rates at the expense of generating more alerts that might be false positives.

Ultimately, this is an organization's decision to balance risk against the resources they have available to investigate the threats. The model evaluation screens contain various charts that make it straightforward for non-expert users to determine the business value that the model will provide. This includes a performance matrix (Figure 12) that summarizes how the model would perform against the training set.



Performance Matrix

	Predicted Non-Event	Predicted Event
Actual Non-Event	Correctly Missed 3833	Incorrectly Hit 176
Actual Event	Incorrectly Missed 17	Correctly Hit 64
	Miss Rate 20.99%	Hit Rate 79.01%

Figure 12. Sample Performance Matrix

The model evaluation page aims to show non-data scientists information about the model that enables them to make smart business decisions. For more advanced users, standard model performance charts (such as, the receiver operating characteristic (ROC) curve, precision-recall (PR) curve, and lift chart) are provided to allow easy comparisons of model versions. These charts are shown in Figure 13.

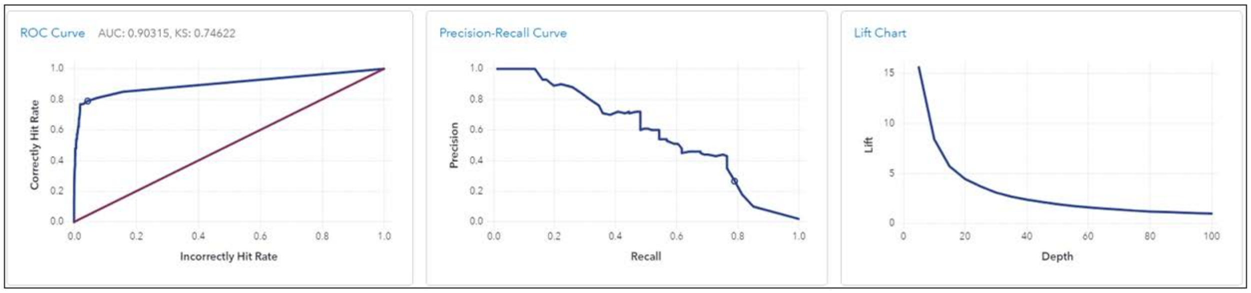
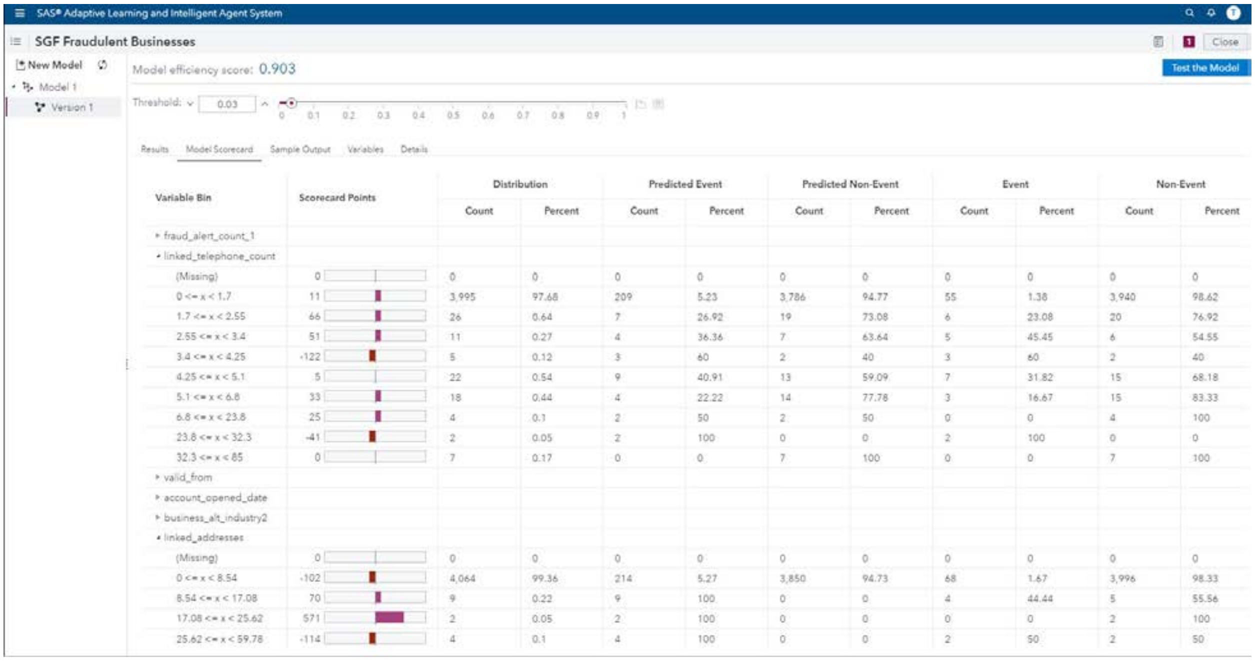


Figure 13. Model Performance Charts

Another key piece of information that is presented is the model scorecard. Since the model is created by the machine rather than the end user, there is a lack of transparency in the model. To alleviate this concern, explainable artificial intelligence (XAI) is required.

The model scorecard screen is the XAI that aids the user in understanding the black box model. This is an independently computed model that serves to help an investigator justify an alert, as well as to help a modeler identify high-risk attributes. The model scorecard is shown in Figure 14.



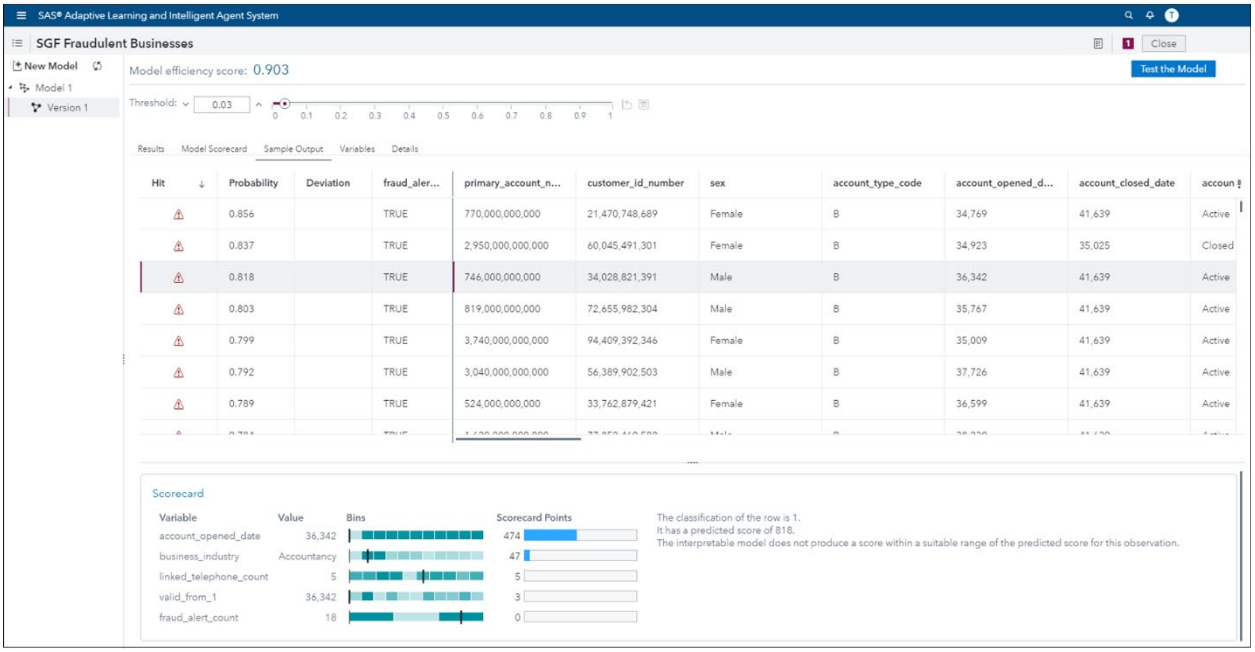
Model efficiency score: 0.903

Threshold: 0.03

Variable Bin	Scorecard Points	Distribution		Predicted Event		Predicted Non-Event		Event		Non-Event	
		Count	Percent	Count	Percent	Count	Percent	Count	Percent	Count	Percent
▸ fraud_alert_count_1											
▴ linked_telephone_count											
(Missing)	0	0	0	0	0	0	0	0	0	0	0
0 <= x < 1.7	11	3,995	97.68	209	5.23	3,786	94.77	55	1.38	3,940	98.62
1.7 <= x < 2.55	66	26	0.64	7	26.92	19	73.08	6	23.08	20	76.92
2.55 <= x < 3.4	51	11	0.27	4	36.36	7	63.64	5	45.45	6	54.55
3.4 <= x < 4.25	-122	5	0.12	3	60	2	40	3	60	2	40
4.25 <= x < 5.1	5	22	0.54	9	40.91	13	59.09	7	31.82	15	68.18
5.1 <= x < 6.8	33	18	0.44	4	22.22	14	77.78	3	16.67	15	83.33
6.8 <= x < 23.8	25	4	0.1	2	50	2	50	0	0	4	100
23.8 <= x < 32.3	-41	2	0.05	2	100	0	0	2	100	0	0
32.3 <= x < 85	0	7	0.17	0	0	7	100	0	0	7	100
▸ valid_from											
▸ account_opened_date											
▸ business_alt_industry2											
▴ linked_addresses											
(Missing)	0	0	0	0	0	0	0	0	0	0	0
0 <= x < 8.54	-102	4,064	99.36	214	5.27	3,850	94.73	68	1.67	3,996	98.33
8.54 <= x < 17.08	70	9	0.22	9	100	0	0	4	44.44	5	55.56
17.08 <= x < 25.62	571	2	0.05	2	100	0	0	0	0	2	100
25.62 <= x < 59.78	-114	4	0.1	4	100	0	0	2	50	2	50

Figure 14. Example Model Scorecard

In addition to the model scorecard, the sample output screen explains the classification of a particular row via statistical analysis. We can select a row of data and immediately see how the model would score it. Also, a scorecard for that row is displayed. This is shown in Figure 15.



Model efficiency score: 0.903

Threshold: 0.03

Hit ↓	Probability	Deviation	fraud_aler...	primary_account_n...	customer_id_number	sex	account_type_code	account_opened_d...	account_closed_date	accoun
⚠	0.856		TRUE	770,000,000,000	21,470,748,689	Female	B	34,769	41,639	Active
⚠	0.837		TRUE	2,950,000,000,000	60,045,491,301	Female	B	34,923	35,025	Closed
⚠	0.818		TRUE	746,000,000,000	34,028,821,391	Male	B	36,342	41,639	Active
⚠	0.803		TRUE	819,000,000,000	72,655,982,304	Male	B	35,767	41,639	Active
⚠	0.799		TRUE	3,740,000,000,000	94,409,392,346	Female	B	35,009	41,639	Active
⚠	0.792		TRUE	3,040,000,000,000	56,389,902,503	Male	B	37,726	41,639	Active
⚠	0.789		TRUE	524,000,000,000	33,762,879,421	Female	B	36,599	41,639	Active

Scorecard

Variable	Value	Scorecard Points
account_opened_date	36,342	474
business_industry	Accountancy	47
linked_telephone_count	5	5
valid_from_1	36,342	3
fraud_alert_count	18	0

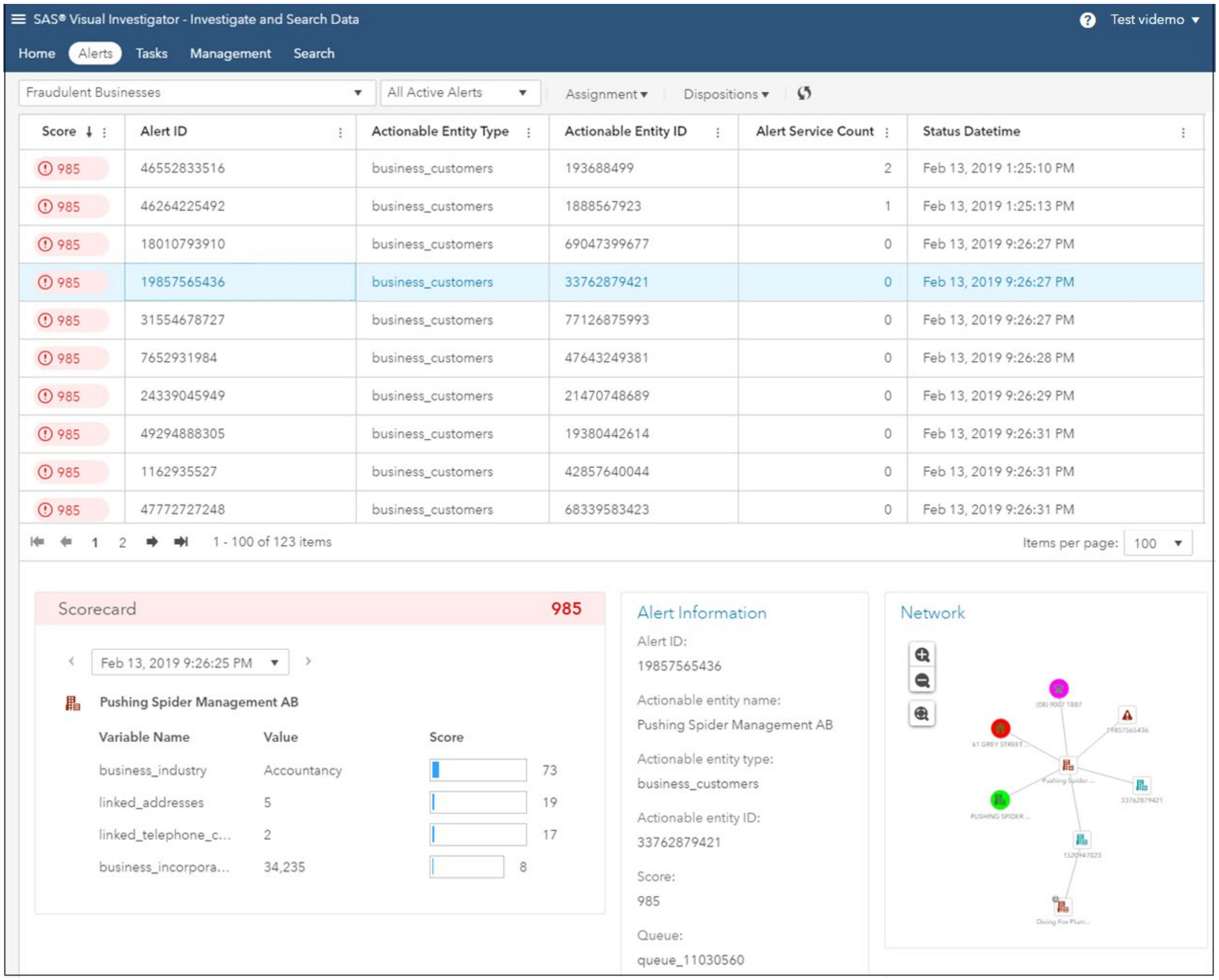
The classification of the row is 1.
It has a predicted score of 818.
The interpretable model does not produce a score within a suitable range of the predicted score for this observation.

Figure 15. Example Scored Row with Scorecard

The information in the scorecard is what is ultimately sent to SAS Visual Investigator and is the basis for analysts starting an investigation.

Once satisfied with the model, we can deploy it to start generating alerts in SAS Visual Investigator. First, new data is scored using the SAS Scoring Engines (stage 4). Based on the threshold that we chose when configuring the agent, alerts are generated in SAS Visual Investigator (stage 5).

Contained in the alert is the explainable scorecard for the associated row of data. This enables us to have a starting point in the investigation. The alert is enriched with the network data surrounding the alerted entity. This is shown in Figure 16.



Score ↓	Alert ID	Actionable Entity Type	Actionable Entity ID	Alert Service Count	Status Datetime
985	46552833516	business_customers	193688499	2	Feb 13, 2019 1:25:10 PM
985	46264225492	business_customers	1888567923	1	Feb 13, 2019 1:25:13 PM
985	18010793910	business_customers	69047399677	0	Feb 13, 2019 9:26:27 PM
985	19857565436	business_customers	33762879421	0	Feb 13, 2019 9:26:27 PM
985	31554678727	business_customers	77126875993	0	Feb 13, 2019 9:26:27 PM
985	7652931984	business_customers	47643249381	0	Feb 13, 2019 9:26:28 PM
985	24339045949	business_customers	21470748689	0	Feb 13, 2019 9:26:29 PM
985	49294888305	business_customers	19380442614	0	Feb 13, 2019 9:26:31 PM
985	1162935527	business_customers	42857640044	0	Feb 13, 2019 9:26:31 PM
985	47772727248	business_customers	68339583423	0	Feb 13, 2019 9:26:31 PM

1 2 1 - 100 of 123 items Items per page: 100 ▾

Scorecard 985

Feb 13, 2019 9:26:25 PM ▾

Pushing Spider Management AB

Variable Name	Value	Score
business_industry	Accountancy	73
linked_addresses	5	19
linked_telephone_c...	2	17
business_incorpora...	34,235	8

Alert Information

Alert ID:
19857565436

Actionable entity name:
Pushing Spider Management AB

Actionable entity type:
business_customers

Actionable entity ID:
33762879421

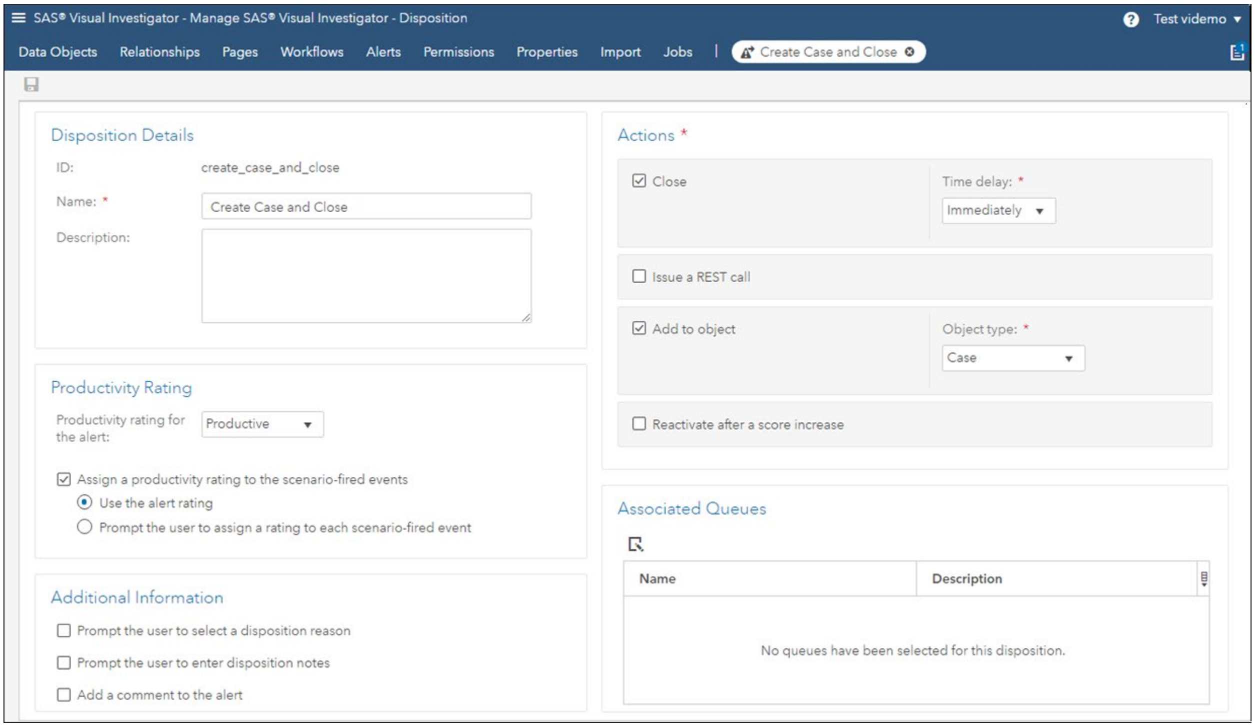
Score:
985

Queue:
queue_11030560

Figure 16. Sample Alert Generated by SAS Adaptive Learning and Intelligent Agent System Surfaced in Visual Investigator

By default, the alerts are prioritized by score to ensure that the most high-risk items are triaged first. The highly visual alerts provide a clear starting point for the investigator's investigation, helping to increase efficiency. The analyst then triages the alert (stage 6) and ultimately dispositions it (stage 7) as either a good alert or a bad alert.

While configuring their solution, administrators can assign disposition settings for analysts. Included in these settings is the disposition productivity rating that includes the options Productive, Unproductive, and Indeterminate. Productive alerts are those that did reveal suspicious activity, unproductive alerts are those that did not reveal suspicious activity (it's a false positive), and an indeterminate alert is one that was inconclusive (for example, put it on hold and revisit in the future). Figure 17 shows the disposition option is “Create Case and Close”, which would label the alert as Productive.



Disposition Details

ID: create_case_and_close

Name: * Create Case and Close

Description:

Productivity Rating

Productivity rating for the alert: Productive

☑ Assign a productivity rating to the scenario-fired events

● Use the alert rating

○ Prompt the user to assign a rating to each scenario-fired event

Additional Information

□ Prompt the user to select a disposition reason

□ Prompt the user to enter disposition notes

□ Add a comment to the alert

Actions *

☑ Close — Time delay: * Immediately

□ Issue a REST call

☑ Add to object — Object type: * Case

□ Reactivate after a score increase

Associated Queues

Name	Description
No queues have been selected for this disposition.	

Figure 17. Sample Disposition Configuration

The intelligent agent is listening to the output of the analyst's investigation and uses these productivity ratings to update its learning table automatically. We can use the job scheduling tools to determine when the intelligent agent should retrain against the updated training data set (stage 1) and create a new model version (stage 2), that can then be evaluated against the previous model (stage 3). At this point, the intelligent agent can continue in an automated fashion learning from the investigations carried out by analysts. If newly created models are better, they can be promoted into production.

Not all organizations will have a labeled data set that enables them to run supervised machine learning. In these circumstances, SAS Scenario Administrator can be used to define expert-driven rules to identify threats. As analysts triage alerts generated from these threats, a history of productive/unproductive dispositions will accumulate. These can be used to create an intelligent agent using supervised machine learning.

There is a second option in which an intelligent agent can be created using unsupervised machine learning techniques to identify anomalies in an organization's data. This approach follows the same adaptive learning process discussed previously. However, whereas at stage 1 there is a training table, in this case the intelligent agent only looks at an unlabeled table. At stage 2 the machine learning algorithms used are different, but the process remains the same. As the system starts generating alerts and they are worked by analysts, the training table is built up. At this stage, the intelligent agent starts to mix both the unsupervised and supervised machine learning techniques to provide a semi-supervised machine learning approach.

At this point the intelligent agent is providing comprehensive detection; the supervised machine learning approach builds up accurate models based on the known previous threats, whereas the unsupervised machine learning approach augments this with the ability to identify new or emerging threats. Likewise, where an intelligent agent starts with a labeled data set, it automatically mixes in unsupervised machine learning. If resources permit, the organization can also use expert-driven detection with SAS Scenario Administrator.

CONCLUSION

Organizations need to identify threats and resolve them as efficiently as possible. However, at a time when they are being asked to cut costs, there are limited resources and limited expertise in data analytics. They are looking to machine learning to fill the gap. SAS Adaptive Learning and Intelligent Agent system is a tool that is valuable to both experts and non-experts. It is a proactive, data-driven detection tool that does not require data science expertise to operate, and can evolve with changing threats. The tool aides in identifying high-risk attributes to investigators, assists in making organizational/business decisions based on risks, and works alongside expert-driven tools in the SAS Visual Investigator environment as part of a comprehensive detection and investigation platform.

REFERENCES

Pozzolo, A. 2015. Adaptive Machine Learning for Credit Card Fraud Detection. Accessed February 14, 2109. Available at http://di.ulb.ac.be/map/adalpozz/pdf/Dalpozzolo2015PhD.pdf.

McKinsey Global Institute. 2016. The Age of Analytics: Competing in a Data-Driven World. Accessed February 14, 2019. Available at https://www.mckinsey.com/~/media/mckinsey/business%20functions/mckinsey%20analytics/our%20insights/the%20age%20of%20analytics%20competing%20in%20a%20data%20driven%20world/mgi-the-age-of-analytics-full-report.ashx.

Karagod, V. 2018. How to Handle Imbalanced Data: An Overview. Accessed February 14, 2019. Available at https://www.datascience.com/blog/imbalanced-data.

ACKNOWLEDGMENTS

I would like to thank Iain Jackson, Jeff Miller, Karlan Wolfkill, Peter Eastwood, Ryan Gillespie, and Tom Mutdosch for their help in reviewing this paper and most importantly for their amazing work on SAS Adaptive Learning and Intelligent Agent System.

CONTACT INFORMATION

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

Rory MacKenzie
SAS Institute Inc.
480 Argyle Street
Glasgow, G2 8NH, United Kingdom
Rory.Mackenzie@sas.com