MACHINE LEARNING INTERPRETABILITY TO AID STRUCTURED ANALYTIC TRADECRAFT

December 2019

Cordis Carter
SAS Institute Inc.
Cordis.Carter@sas.com

Evelyn Fox
SAS Institute Inc.
Evelyn.Fox@sas.com

Contents

Abstract

This research effort focused on introducing machine learning interpretability techniques to aid structured analytic tradecraft. Analysts are hesitant to use complex mathematical models in their analysis due to difficulty in interpreting and communicating the results. A hands-on workshop was conducted introducing various model interpretability techniques to understand (1) how analysts might employ model interpretability methods and (2) what methods and techniques can increase confidence in communicating complex modeling results. A scenario and machine learning model pipeline were developed with two publicly available datasets (*The GDELT Project* and the *World Food Programme*) to provide the analysts a representative problem set that they may be tasked with in their current analytics workflows. Analysts with more technical backgrounds, such as in math or computer science, found interpretability techniques applied throughout the entire data analytics lifecycle as critical in the analysis process; analysts with limited mathematical knowledge struggled to understand how a variety of interpretability techniques could support or contradict their analysis. Despite the variation in backgrounds, common feedback among the analysts was centered around the lack of understanding as to what different types of models do, how they are used, and what types and quality of data are required. Ultimately, they agreed that rapid adoption of advanced analytics is more likely in an area where there is low risk in the incorrect application of complex modeling algorithms.

Introduction

Background

The intelligence analyst today increasingly has access to new and relevant data sources continually being developed, providing more methods and data sources to cull through to make assessments and accomplish their requirements. But every time a new data source is made available and the cost of technology decreases to store more and more data, analysts are expected to analyze that data to find the points they need to make assessments. In 2005, an analyst may have had to read through 30 text messages to see if any contain intelligence value. In 2019, that same analyst may be expected to go through thousands of text and chat messages, correlating with social media posts and other public data, to perform the same analysis.

It has become evident that without taking advantage of those same advances in technology to process the growth in size and complexity, an ever-growing backlog of data will either never be analyzed or will be analyzed so late that the resulting assessments are less significant. Like any new capability or technology, there are challenges with incorporating this capability into existing workflows. These challenges include the lack of trained resources, at all levels of the spectrum, to fully adopt these tools broadly across the enterprise. Therefore, the research team decided to conduct a study not where the analysts were forced to manipulate or process data, build models and/or workflows, but rather focus on what they understand, what they know, and what they can deduce from the data and modeling results. The desired outcome of the study is to determine what works for the analyst community in building trust, understanding, and ultimately incorporating these advanced methods into their day-to-day analytic process.

Why Model Interpretability?

With the widespread collection of data and expanding computational resources, machine learning is becoming more prevalent in daily life and applied to more complex problems. As this problem set evolves and higher-risk problems are considered, the output and process to which the model decides on an outcome becomes increasingly important. Take for instance the application of machine learning to autonomous vehicles and their identification of pedestrians walking across the street. If the model found that the most important feature to make this decision is the appearance of crosswalk lines, then this would not cover the case of a pedestrian jaywalking. Since hitting a pedestrian is a very high-risk outcome, it becomes increasingly important to incorporate model interpretability methods to help understand the potential bias in the model. It is not necessary for consumers of this information to understand the technical and mathematical details of the modeling algorithm, but it is critical to be able to interpret and understand how the model correctly identifies pedestrians in terms understandable to decision-makers that need to appropriately manage risk before employing these novel AI techniques.

> *"Interpreting is essential to our futures. This is the way the world is going. It should be brought to the masses"*
>
> *- Intelligence Analyst*

Goal

In order to leverage recent innovations in machine learning and automation on the increasing volume and variety of available data, analysts need to be able to interpret and understand the results of these complex algorithms and explicitly document how their intelligence assessments were reached. The primary outcome is to determine which techniques analysts prefer and are willing to adopt into their tradecraft by learning and discussing the usability, practicality, and confidence in using model interpretability methods, graphs, and visuals to provide defensible data-driven evidence to decision-makers. The secondary outcome is to learn how an organization can incorporate these kinds of interpretability methods if the organization does not have a history of using them nor a depth of trained personnel who understand these concepts.

Literature Review

As analytic methods become more complex with the emergence of machine learning and artificial intelligence, the more difficult it has become to understand the results and underlying methods for those without an advanced technical background. Due to this, the LAS and SAS teams collaborated in a literature review to detail various novel techniques that can be used to explain the results of complex machine learning algorithms to determine which methods might be highlighted during the analyst workshop.

Explainability vs. Interpretability

The main topic of discussion was industry's interchangeability around the definitions of model interpretability versus explainability. As the project developed, it became clear that from an intelligence analyst perspective, there is a significant difference between understanding a model, and explaining how a model works from a technical and mathematical perspective. For the purpose of the study, the team formally defined these terms to ensure a common definition to move forward with. Based on the extensive literature review, the teams agreed on the following definitions:

Interpretability: the extent to which analytic results are understandable and discernable using a vocabulary that is meaningful to the consumer of the output. The ability for someone to comprehend and explain to others why a certain decision or prediction has been made.

Explainability: the extent to which the details of a machine learning workflow can be described in an accurate and understandable way. The ability to understand the inner workings and completeness of how the model works from a mathematical perspective.

The team decided that the workshop would focus on model interpretability rather than explainability. The ability to understand model results and communicate them effectively to decision-makers was determined more practical for analysts as opposed to having a full understanding of the inner workings of machine learning models. The workshop and project were developed with this intent in mind.

Proposed Scenario and Model Development

Scenario

A scenario was developed to provide analysts a relevant problem set that they may be tasked to perform in their current analytic process. At the heart of the scenario was an intelligence question, created to provide an environment in which analysts could look at data and advanced analytic models, and focus on the analytic process without getting entangled in the details and validity of the specific intelligence question. The intelligence question was: *"Can we use historic news reporting and economic data to predict violence in Afghanistan?"*

There is an abundance of reporting and known dates around specific events, such as elections, in Afghanistan over the last 18 years. In looking at a known location like Afghanistan, there is the ability to look at data in a variety of ways to simulate the daily intelligence workflow and assessment requirements that an analyst could face. This example allows for the incorporation of other data, such as economic data, to be overlaid on the event data to simulate the use of multiple data streams to perform analysis on. Additionally, this is a topic that was relatively simplistic for analysts to understand. Even if they are not working the Afghanistan/Pakistan problem set daily, there would still be a general familiarity around topics related to Afghanistan and violence. This was crucial to ensure that analysts would not decline to participate in the discussion due to lack of understanding and comfort around this topic. As one participant said, "[there is a] difference in working with a data set I know versus a new one I don't know."

> *"I want to know what data was used in an advanced analytics process. Test it on something that you know is right, and verify it"*
>
> *- Intelligence Analyst*

Data Sources

Two data sets were cleansed and provided to the analysts from *The GDELT Project* and *World Food Programme*. The GDELT Project "monitors the world's broadcast, print, and web news from nearly every corner of every country in over 100 languages," capturing features such as known organizations, locations, and sources. Each day has a publicly accessible data file that captures these features dating back to April

2013. Data was pulled from April 2013 through January 2018 and appended into one data set for cleansing and standardization prior to it being analyzed.

The cleansing process consisted of creating a data subset, deduplicating, and enriching. As mentioned before, The GDELT Project has their own process for capturing certain features about each article, including which countries are involved or mentioned. The team used these features to subset the original dataset down to articles that The GDELT Project determined were related to Afghanistan to narrow down the volume of data that would be relevant for the proposed scenario. Once the subset was created, the data was then deduplicated by the URL provided to ensure there were not multiple occurrences of the same article. Enrichment of The GDELT Project data included retrieving the article URL, creating and deploying a Python script to scrape the raw textual content from the article, and flagging the text based on occurrence of specific topics of interest.

SAS Visual Text Analytics was applied to the raw article text to flag articles for specific topics in order to enrich the text data for predictive purposes. Both user-defined and natural language processing (NLP) techniques were used to create various rules to identify topics that might be indicative of increasing violence in Afghanistan. Table 1 displays the nine text topics that were created, a description, and it's use in the model development process. Text topics are not mutually exclusive; therefore, an article can be flagged for containing content related to multiple topics.

Text Topic	Definition	Role
Afghan Election	Binary = 1 if an article contains content related to Afghan elections	Model Input
Afghanistan	Binary = 1 if an article contains content related to Afghanistan in general	Model Input
Incident	Binary = 1 if an article contains content related to violent incidents	Model Input
Increased Risk	Binary = 1 if an article contains content to increasing risk	Model Input
Increased Violence in Afghanistan	Binary = 1 if an article contains content related to increasing violence specific to Afghanistan	Development of Target Variable
Increasing Violence	Binary = 1 if an article contains content related to increasing violence	Model Input
Political Related Violence	Binary = 1 if an article contains content to political related violence	Model Input
Protest	Binary = 1 if an article contains content related to protests	Model Input
Threat	Indicator (0, 1, or 2) to group articles into a hierarchy of threat levels based on GDELT's CAMEO event codes (higher = more threatening)	Development of Target Variable

Table 1. *Text topics created with SAS Visual Text Analytics to enrich The GDELT Project's article data*

The target variable, AFG_INCREASED_VIOLENCE, was developed through a combination of the text topics *Increased Violence in Afghanistan* and *Threat*. If an article was flagged for containing content related to increasing violence in Afghanistan and had a threat level of 1 or 2, then the article was assigned a value of 1 for the binary target AFG_INCREASED_VIOLENCE. This process allowed the SAS team to develop a supervised predictive model pipeline for the analysts to work through.

Economic data was provided from the *World Food Programme*. Features included prices for various commodities at the monthly level (bread, fuel, sheep, wheat, exchange rate). Additional features were generated to look at the percent change in the price of each commodity during to the month the article was published compared to the previous month. Percent change was used in order to keep the prices for each commodity on a standardized scale of pricing. A full list of the variables used for model development meant and exploration is provided in the Appendix.

Model Development

A machine learning model pipeline was developed in SAS Visual Data Mining and Machine Learning (VDMML) to determine the best performing model in predicting violence in Afghanistan based on news articles and economic indicators. Multiple machine learning models were compared head-to-head as shown in Figure 1. The gradient boosting model was automatically chosen as the best-performing (champion) model based on specific model fit statistics criteria, in this case misclassification rate.

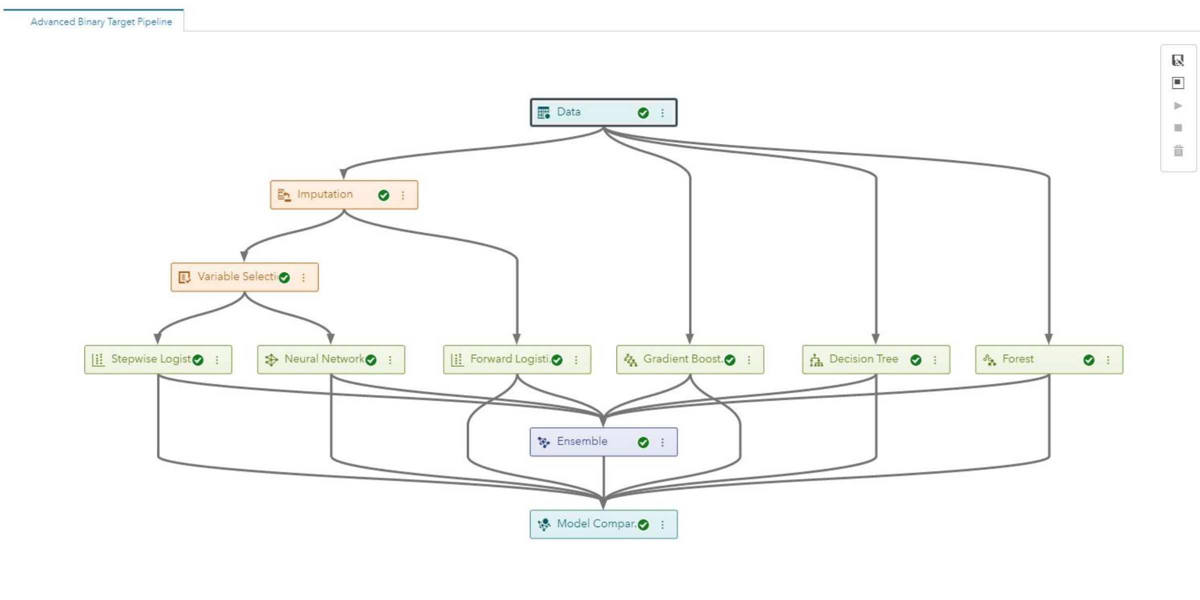


Figure 1. *Advanced machine learning model pipeline to predict violent events in Afghanistan*

The purpose of building the model was to provide a representative example of an advanced machine learning model that analysts could discuss, validate, and interpret, not for the purpose of developing the most robust model that is production-ready.

Analyst Workshop

Overview

The team hosted a one-day workshop with analysts from different analyst backgrounds, with a wide range of technical proficiency, to determine how an analyst would go about incorporating advanced analytics and machine learning techniques into their existing workflows, and what would be required to do so. The research team facilitated two breakout sessions to: 1) capture analyst feedback as they were walking through a data analytics workflow to answer an intelligence question, and 2) to understand what information an analyst would need to interpret and explain model results to their peers and leadership when given an unstructured scenario.

The five-hour workshop was held on November 13, 2019, at the Poulton Innovation Center on the North Carolina State University campus. There was a total of nine analysts accompanied by four LAS personnel that were participants in the workshop. The analysts were briefed on the project overview, goals, and instructions for the two interactive sessions. The participants were split into three groups for the interactive activities, each group guided by a SAS analytics expert.

Hands-On Activity

The SAS team developed a hands-on activity for the analysts to work through in groups during the workshop. The activity, presented using SAS Visual Analytics, consisted of 11 exercises detailing the analytic workflow that an analyst might use to incorporate these machine learning methods into their tradecraft. The first seven exercises guided the groups through feature descriptions, data profiles, correlation matrices, and visual exploration of both the article and economic data. Figure 2 highlights two of these visuals that depict the density of article collection over time using a heat map, as well as the top 15 sources where articles were pulled from. A diverse set of visuals were provided to the analyst throughout the exercises in order to collect feedback on usability of communicating complex structured and unstructured data in effective ways.

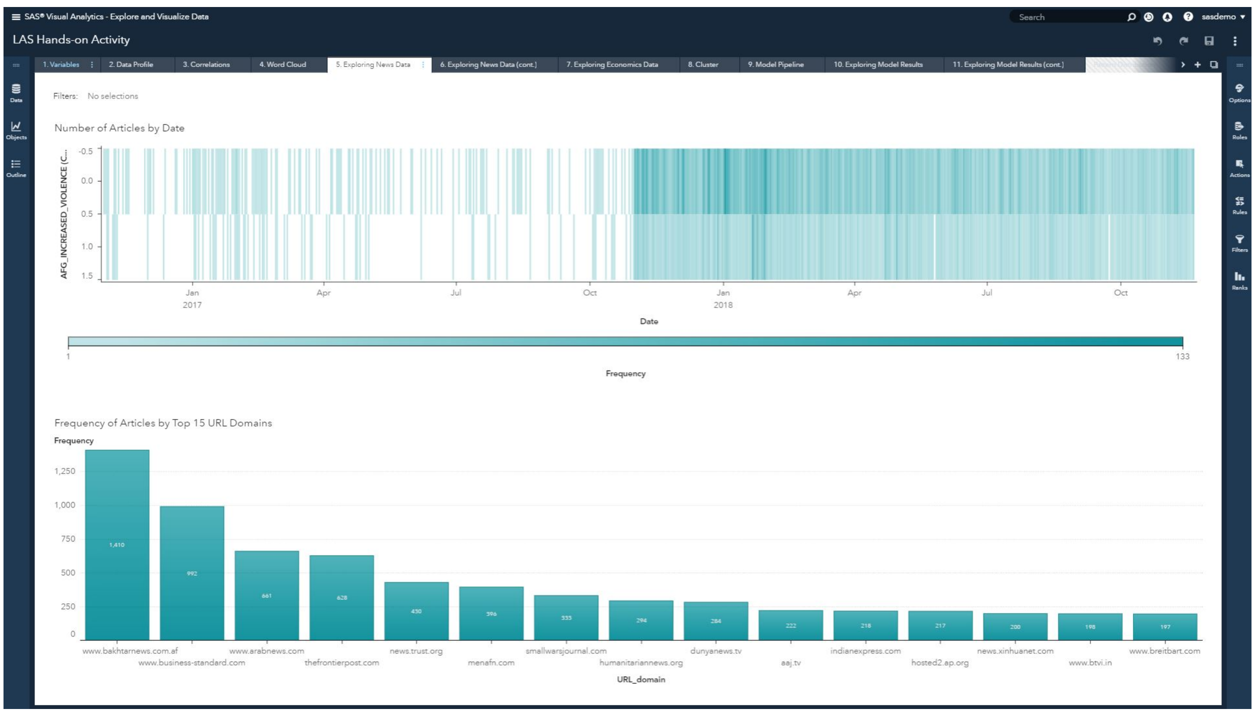


Figure 2. *Analyzing coverage trends in source bias and article volume over time*

The second half of the hands-on activity focused on different modeling methods and interpreting the results. Analysts were provided the results from the best-performing (champion) model, the gradient boosting model, as displayed in the pipeline depicted in Figure 1. The exercises included calculating various performance metrics from a confusion matrix, as well as analyzing the partial dependence plots of the top 10 influential variables. A confusion matrix, shown in Figure 3, “describes the performance of a classification model on a set of test data for which the true values are known” (Data School).

n=165	**Predicted: NO**	**Predicted: YES**
Actual: NO	50	10
Actual: YES	5	100

Figure 3. *Sample confusion matrix (Data School)*

Analysts were asked to calculate different performance metrics such as accuracy, misclassification rate, false positive rate, and others. This method was presented to the analysts first, as it is often used as a simplistic way to grasp an understanding of model performance without the need of a technical background.

Analysts were introduced to partial dependence plots for the top 10 influential variables in the gradient boosting model. “Partial dependence plots show the marginal effect one (2D) or two (3D) features have on the predicted outcome of a machine learning model” (Molnar). While more computationally technical, partial dependence plots can be helpful in quickly determining the relationship between the target and a feature as linear, monotonous, or more complex. Figure 4 displays two partial dependence plots that were presented to the analysts, showing one example of an interval variable (month) and one categorical variable (violence text topic).

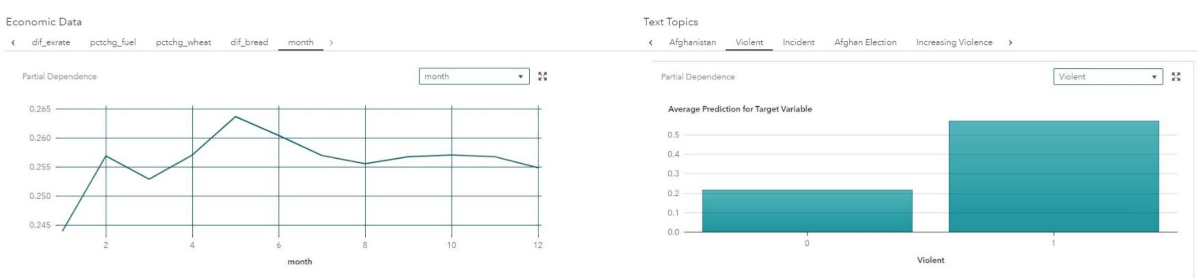


Figure 4. *Partial dependence plots were introduced to assess analyst reaction to this model interpretability technique*

The participants were given two hours to work through the hands-on exercises within the three groups. A SAS analytics expert accompanied each group in order to facilitate conversation and answer any technical questions that might arise during the activity. All groups came together at the end of the session to participate in an open discussion about best practices, usability, and applicability of the content they worked through and how that compared to their own analytic processes.

Analyst Storyboard Activity

The afternoon activity took the analysts from exploring the data and understanding models and how to interpret them, to a theoretical exercise in implementing models and their results into a workflow. Many analysts need to have the opportunity to practice with a new tool or process to build confidence that it will either augment or replace their existing methodology. This hypothetical exercise gave the groups the time to talk through and decide on a scenario, then take elements from the morning session, including visuals and lessons learned, and apply them to the scenario.

The basis of the scenario was as follows. As an analyst who has models and modeling results at their disposal, an analytical tasker requires an assessment where the analyst must use model results in order to provide their assessment to a senior analyst or decision-maker. What type of data points, graphs, visuals, or other data-driven evidence provides the strongest support in proving their assessment? The workshop facilitators provided two scenarios, while allowing each group the option to utilize their own scenario if desired.

> *"As a beginner, I need something simple to start. Start smart with understanding. I would need someone I respect to trust the results before I would fully adopt"*
>
> *- Intelligence Analyst*

The first scenario presented was using social media analysis to assist in operations planning with the least negative impact across the population. The analyst is working in direct support of a military unit trying to determine a series of future operations targeting an adversary. While working through the operational planning cycle, leadership decides that they want to maximize effects against the enemy, while minimizing the negative perception of the United States among the local populace. The question becomes: which has a higher impact on public opinion, drone strikes or raids? Using a historical scrape of social media platforms used in the country of operations, natural language processing is applied to the data to quantify sentiment of the postings, responses, and re-postings. The analysis is then run against new data as it becomes available to assess the impact of the event and responses of the population following drone strikes and raids. The facilitator drew out a mock storyboard where he had visualizations about the density of data collection over time, the changes in overall sentiment over time, and occurrence of key events. Using a combination of the storyboard elements, the facilitator demonstrated how courses of action could be assessed in minimizing the increase in negative sentiment.

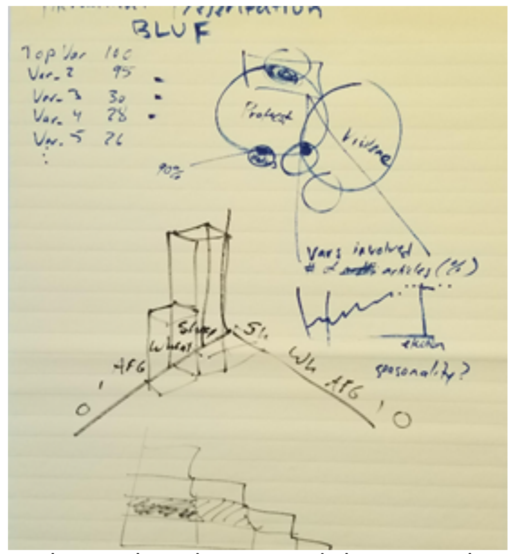
> *"Trust, but verify the model while continuing to use what I'm already doing to make sure the model is producing the correct results"*
>
> *- Intelligence Analyst*

The second scenario the analysts had to choose from consisted of the scenario and data used in the first session, the Hands-On Activity. If a senior analyst or supervisor asked an analyst to assess the occurrence of violence in Afghanistan during an upcoming election, what would they need to see within the data, and what visualizations would they take to leadership in order to support their assessments? The analysts worked in a small group to discuss the data and visualizations that were examined in the first session to determine what they would use, change, or discard to support their assessment.

The final option for analysts was to come up with their own unclassified scenario. The scenario could be something they encounter in their daily responsibilities or challenges that commonly occur. The goal was to pick a scenario that would lend itself to characterizing how they envision understanding and interpreting the results of advanced analytics outputs. Each group was then asked to walk through what type of data and visualizations would be required, and what sort of knowledge they would need in order to support their data-driven assessment.

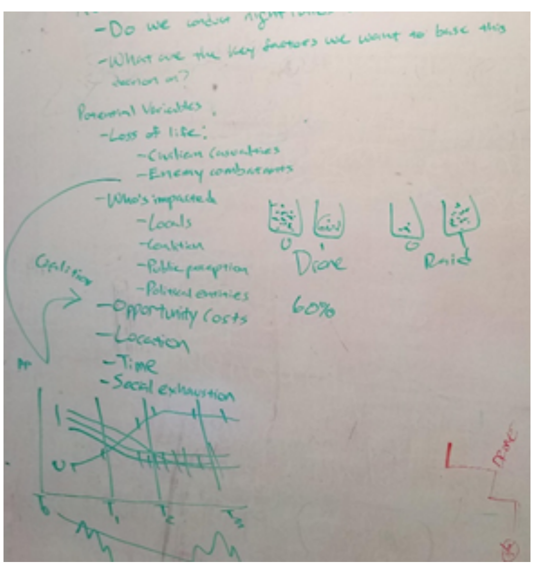
Of the three working groups, coincidentally, each group picked a unique scenario option that was provided. A summary of each group's discussions is detailed below.

Group 1



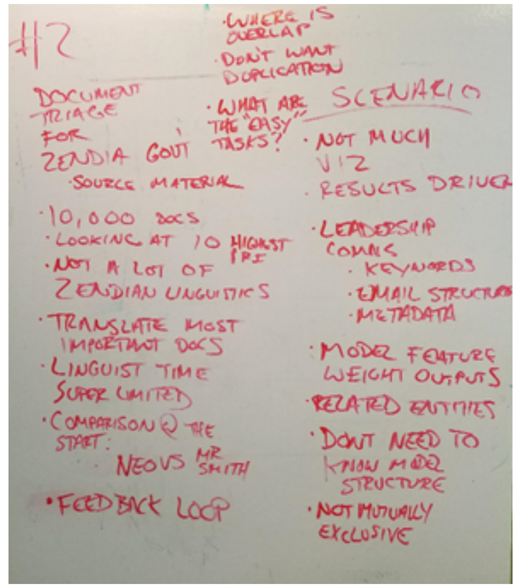
This group picked the scenario where they worked with the data from the first session, the Hands-On Activity. In looking at how they presented a prediction of violence, Group 1 believed that listing the importance of the top variables driving the model results would depend on the specific audience – whether that be a senior analyst or a commander. Group 1 walked through which visuals would work best and discussed how they could visually represent the most important variables. Group 1 considered a 3D visualization to show correlations between variables, while also using 'shading' to demonstrate the level of dependency between the variables, with an ability to hover over the visual to show further contextual information. There was no consensus among the group if this was a suitable visual to assist in interpretability. Finally, since they would be doing predictive analytics, they determined that a timeline representation would be critical in their presentation of an assessment.

Group 2



This group chose the scenario related to measuring impact of population sentiment based on drone attacks versus night raids. They began their discussion by assessing which population groups they would want to assess the sentiment analysis on: do we want to assess public reaction? Do we want to assess the intent of local population? Do we want to assess reaction in the United States to a series of offensive operations? Do we want to assess reaction of partner nations to possible courses of action? Group 2 wanted to devise a method to weigh the variables in the data and look at them as a connected time series. By doing so, the group looked to gauge a level of ambivalence within the reactions, and then use a timeline visualization to show the frequency and level of reaction over time. Next, the group decided that a (logical) decision tree would be helpful as a part of the decision process to visually represent the decision paths of the various courses of action (night raids vs. drone strikes). The goal was to show what data and variables led to the final recommendation. The group stated that a data-driven decision should be based on a series of models that ultimately inform and provide evidence to the human decision.

Group 3



This group decided to create their own scenario, in which an analyst has been provided 10,000 documents, and he or she only has the time to read and analyze 10 documents per day. They pondered questions such as: "What would an analyst like to see?", "How are the 10 documents per day selected?", "How critical is it that the top 10 documents are the best ones to analyze?", "How accurate do we need to be in the prioritization and 'triaging' of documents?". Group 3 had an analogy of having multiple Internet search engines on multiple tabs searching on the same terms and looking at the results to compare. The results would determine the validation and trust in the model and methodology used to select those 10 documents. Group 3 stated there should be multiple methods to select the highest priority documents, with feedback loops between them all.

Results

Overview

Machine learning and modeling is becoming more prevalent across the intelligence realm, but analysts do not agree on how to best implement it, where it should be implemented, and what level of training and ability do analysts need to incorporate it into their workflows. The analysts at the workshop all understand the expansion in data size, data sources, variety of data, and the challenges that come with it. They understand that the old, tried-and-true methods will not be able to keep up without advanced analytic methods to tackle this growth. "Government leadership needs to convince the analyst community" is an example of how one analyst thinks incorporation of these methods will have to be imposed to be adopted. In order to understand the value of this workshop, it is important to look at both the common feedback, and variation in feedback, in order to look at ways to help analysts build these processes into their workflows faster, more effectively, and more efficiently. During the workshop, analysts, regardless of technical prowess, developed a better understanding of each other's perspectives and challenges. More engagement of this type will lead to better collaboration and partnerships.

> *"We need to manage expectations about what machine learning can do. Perhaps provide an introduction to machine learning, but don't get too in the weeds. ML and AI are buzzwords right now"*
>
> *- Intelligence Analyst*

Common Feedback Among Analysts

- Regardless of an analyst's level of understanding and familiarity of data science, there are still knowledge gaps when it comes to models, algorithms, and adopting these techniques. It is the level of comfort and familiarity that differs, and how to solve the issue of blending historical analyst methodologies with advanced analytics and machine learning techniques.
- There is a lack of understanding among most analysts as to what different types of models do, how they are used, and what types of data are required.
- Embedding a small number of data science professionals in various analyst teams would enable the proper application of advanced mathematical methods, while analysts can remain focused on their analysis expertise. They agreed that, over time, analysts will become more proficient at understanding and interpreting analytics outputs, but there will likely always be a technical gap in understanding how advanced analytics can be applied that will need to be filled by an expert.

> *"As an analyst, I want to be able to trust [the analytic outcome]. I don't have the time to try to understand it. There should be someone who does this for me. I don't want to learn another job."*
> *- Intelligence Analyst*

- The significance of the analysis and the importance of the outcomes plays a major role in trusting the incorporation of models into existing workflows; the graver the decision, the less likely to trust replacing manual processes. The rapid adoption of advanced analytics is more likely in an area where there is low-risk of an incorrect application of advanced analytics; for example, using text analytics to automatically prioritize and 'triage' large volumes of unstructured documents for analyst review. A higher-risk scenario, such as utilizing an advanced mathematical model to action a target, will take much longer to build trust in the process.

> *"Analysis directly supporting the warfighter is more critical than long-term economic analysis. A place to start the adoption could be in support of economic or financial analysis."*
> *- Intelligence Analyst*

- The user experience is very important. That goes all the way down to the labeling of visuals, which was identified as very important. If there is not a clear uniformity of colors, that will lead to hasty inaccurate assessments. Too much information, and information that is not understandable, can be detrimental to building trust and confidence in models. For example, if the labels and numbers along one axis of a chart are not clear in their meaning, it could be better to leave the axis unlabeled. Visuals need to be standardized – if the format differs, it will lead to missed assessments. Visuals are only good if they explain the data in the clearest possible way.

> *"Predicting versus triaging information is different. Adoption may occur quicker when starting with analysis with a 'lesser' impact or risk. This may be an easier place to start."*
> *- Intelligence Analyst*

- Trust in the data, and trust in the workflow, is paramount to an analyst's desire to incorporate models and advanced analytics into their process. Analysts are about what tools they use and do not use, and a lot of that is dependent upon the results they get, and how easy it is to achieve their outcome. In order to bring a new tool or technique into an existing workflow, or even replace existing steps, there must be an inherent trust in that technique or tool to adopt it. But there needs to be a better understanding between the analyst, the data scientist, and the machine learning techniques available to effectively incorporate into the process.

> *"Trust, but verify the model while continuing to use what I'm already doing to make sure the model is producing the correct results."*
> *- Intelligence Analyst*

Variation in Feedback Among Analysts

- Analysts with more technical backgrounds, such as in math or computer science, find interpretability techniques applied throughout the entire data analytics lifecycle as critical in the analysis process; analysts with limited mathematical knowledge struggle to understand how a variety of interpretability techniques could support or contradict their analysis.
- Some analysts think that knowledge and training on machine learning and AI to improve their analytic workflows will be required in the future; other analysts do not think they will be required to know these skills in the future. Some analysts are very interested in learning the details, others did not care about ever knowing how the math is done, just give them the results, so long as the results are correct, of course.

References

Al, Zelros. "A Brief History of Machine Learning Models Explainability." 24 Sep 2018. https://medium.com/@Zelros/a-brief-history-of-machine-learning-models-explainability-f1c3301be9dc.

Google Jigsaw. *The GDELT Project.* www.gdeltproject.org/data.html.

Hall, Patrick and Gill, Navdeep. "An Introduction to Machine Learning Interpretability: An Applied Perspective on Fairness, Accountability, Transparency, and Explainable AI." April 2018. https://pages.dataiku.com/hubfs/ML-interperatability.pdf.

Kaynar Kabul, Ilknur. "Interpretability is crucial for trusting AI and machine learning." *SAS Blogs,* 18 Dec 2017. https://blogs.sas.com/content/tag/ai-interpretability/.

Lipton, Zachary. "The mythos of model interpretability." *acmqueue.* May 2018.

Molnar, Christoph. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.* 10 Mar 2019. https://christophm.github.io/interpretable-ml-book/.

Preece, Alun. "Asking 'Why' in AI: Explainability of intelligence systems – perspectives and challenges." *Intell Sys Acc Fin Mgmt.* 29 Jan 2018. https://doi.org/10.1002/isaf.1422.

"Simple guide to confusion matrix terminology." *Dataschool.io*, 25 Mar 2014. www.dataschool.io/simple-guide-to-confusion-matrix-terminology/.

Vellido, Alfredo. "The importance of interpretability and visualization in machine learning applications." *Neural Computing and Applications,* 23 Jan 2019.

World Food Programme. Afghanistan – Food and Commodity Prices. https://data.humdata.org/dataset/wfp-food-prices.

Appendix

Variable List

Actor1CountryCode	3-letter Country code for which country is determined Actor1 per GDELT
AFG_INCREASED_VIOLENCE (target)	Indicator Variable (0 or 1) if an article is tagged for increased violence based on GDELT event codes
Afghan Election ***(text category 1)***	Indicator Variable if an article contains content related to Afghan elections
Afghanistan *(text category 4)*	Indicator Variable if an article contains content related to AFG as a whole
Date	Date the article was published
Day	Day of the month the article was published
Dif_(commodity)	The difference between the current price of a commodity (Bread, exchange rate, fuel, sheep, or wheat) and the previous month's price of that commodity
Exrate_price	The exchange rate during the month the article was published
ID	Article ID provided by GDELT
Incident *(text category 3)*	Indicator variable if an article contains content related to violent incidents
Increased Risk *(text category 2)*	Indicator variable if an article contains content related to increasing risk
Increased Violence in Afghanistan *(text category 11)*	Indicator variable if an article contains content related to increased violence specific to Afghanistan
Increasing Violence *(text category 9)*	Indicator variable if an article contains content related to increasing violence
Lag_(commodity)	The price of the commodity a month before the article was published
Month	The month the article was published
Pctchg_(commodity)	The percent change between the current month's and the previous month's price of the commodity
Political related Violence *(text category 7)*	Indicator variable if an article contains content related to political related violence
Protest *(text category 8)*	Indicator variable if an article contains content related to protests
Text	Article Text
Threat	Indicator variable (0,1, or 2) to group articles into different threat levels based on GDELT event codes (higher=more threatening)
Threat/0 *(text category 5)*	Indicator variable if an article contains content related to level 0 threats
Threat/1 *(text category 6)*	Indicator variable if an article contains content related to level 1 threats
Threat/2 *(text category 10)*	Indicator variable if an article contains content related to level 2 threats
uniqueid	Automatically generated unique ID
URL	Full URL where the article was published
URL_domain	URL domain (general website) where the article was published
Violent	Indicator variable if an article contains content involving violence based on GDELT event codes
Year	The year the article was published