Subject: Feedback for NISTIR 8312
Date: Thursday, August 20, 2020 at 10:49:55 AM Eastern Daylight Time
From: Brennan Murphy
To: NIST Explainable AI
Attachments: image.png, image.png

Hi,

I wanted to offer some feedback on NISTIR 8312.

1) I could not easily and quickly glean what the purpose of this report was intended to be. Are you developing standards through which organizations can evaluate xAI claims made by companies offering such services? Is the intent to create standards that serve a technological purpose, for example, an open standard for xAI that all companies should write code to in order to win the stamp of approval? Is some technical integration issue driving the need? It just isn't real clear at the outset what the goal is. I don't get a strong sense of who is the intended audience for this work---data scientists, industry folks, etc? This should also be very up front in the work. I would recommend targeting the less technical audience from a least common denominator perspective. If you're writing about xAI, the work itself should be widely intelligible in my opinion.

2) As you've captured well, the scientific literature on xAI is all over the map. There is not widespread agreement on a particular set of concepts being the gold standard when it comes to explaining what xAI is all about.

However, I feel very strongly about the starting point you have taken with this work. If I can speak broadly about the entire paper, I would say you have taken a model-centric approach. Thus, you have characterized various nuanced issues that arise for explainability based on which model is under consideration. This gets very entangled in complex data science very quickly and renders the overall work less intelligible to the wider potential audience. I would even argue that if the target is a technical audience, this work as it is now is a bit too convoluted. (no offense) It is choppy as a result--doesn't flow well.

Instead, what I propose is that you take a very data centric approach to this work. If you do that, you have a full set of conceptual resources to draw upon to structure a narrative around how the problem of xAI arises. I'm appending a simple graphic we use to discuss analytics through a step progression. See attached step chart below.

AI and modeling generally speaking are subsets of data analytics. The value of analytics broadly speaking has to do with what tense the analytic provides--past and present analytics are valuable to know what happened and why but future tense analytics are more valuable since you can correct course and navigate around future problems.

The key concept, in my humble opinion, that is missing from your work here is that of data dimensionality. My company works with genomics data which has some of the highest number of variables of any data sets. xAI becomes a challenge as dimensionality begins to exceed what a human can handle. When a model begins to treat 25 to 50 variables or more as part of a target prediction, being able to convey how those variables operate together and relate to that target variable becomes increasingly difficult to comprehend. This is a key concept that needs to be better developed in your paper.

Once that groundwork is laid, you can begin to introduce the complexities that arise from algorithms and models. Linear and logistic regression are models designed to work with lower dimensionality data involving only a few variables. Deep neural network and other variants are designed to work with higher dimensionality. Our technology, Topological Data Analysis is designed to work with ultra-high dimensionality for example stemming from complex phenomena like genomics data or stemming from complex phenomena like manufacturing steps of computer parts like hard drives or RAM or factors impacting a large financial institution's liquidity or risk.

As you stated, in the financial sector, the key is for models to be explainable to regulators. But you have to be clearer

in your paper that what youre ultimately explaining to a regulator is the risk your bank is taking with its loan portfolio for example. That's the underlying phenomena under analysis when the issue of xAI arises. Once you create an AI model, you have to be able to say why an algorithm was chosen and this usually refers back to the data and specifically the *dimensionality* of that data. For example, you might say, "we had loans failing 8 % of the time in our particular market. We weren't aware of what the key factors were within those failures so we modeled out and learned that factor 1, 2 and 3 appear together in 80% of the cases. Therefore, our model surfaces these factors for our consideration of any decision based on our data that meets that criteria." Context is key in explainability. Data centricity allows better context.

Models are based on data that describes a physical or tangible reality. That foundation does not come out easily in your paper in its current state. The relationship of a model to reality is mediated through data conceptually speaking. Ultimately, AI is about prediction, ie, the future of some phenomena of interest.

In general, I feel it is really important to always understand that data science is science. Science is about looking for patterns in reality and being able to predict the behavior or future of some thing or system or object. Data science is the same except we're looking for patterns in data that reflects the patterns found in reality. If you don't drive home these key concepts, you'll be floating around in the clouds and non-practitioners won't get much from your paper due to the steep learning curve to grasp many of the concepts you deploy.

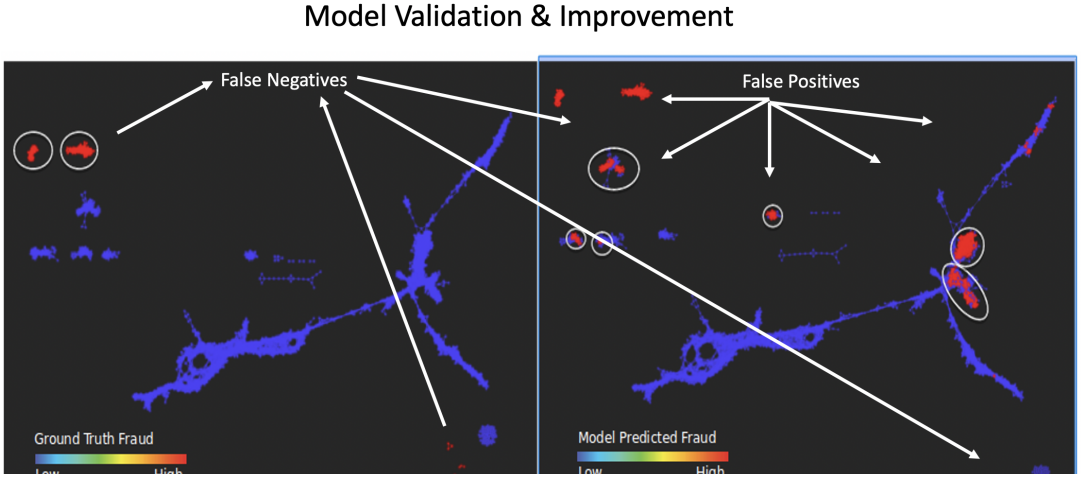
3) xAI has benefits to data science practitioners. xAI makes it possible to model faster, with more accuracy and less sustainment. Therefore, xAI is an inherent efficiency for practitioners. yes, it also has increased security benefits since when you know how a model works, you can assess what is required for its security.

4) I'm not sure I like the names of the 4 principles. What about this scheme instead:

Explanation ------> Objective Explanation
Meaningful--------> Subjective Relevance
Explanation Accuracy -----> Traceability or Forensic Traceability
Knowledge Limits -----> Context Appropriate

5) Although you use the word intelligibility in the paper, I think the concept itself warrants further inclusion in your work. IMO, the discipline of philosophy provides the clearest representation of what intelligibility is. Here is an example from Alasdair MacIntyre:
https://link.springer.com/chapter/10.1007/978-94-009-4362-9_4

6) There's an aspect of xAI that is missing from this paper. In simple low dimensional business analytics or business intelligence, a few variables are presented in a simple chart or graph. This is usually heavily time series oriented. But what about graphics for high dimensionalty data? This is what leads to global comprehension of a complex dataset. I'm going to give you a TDA based example:



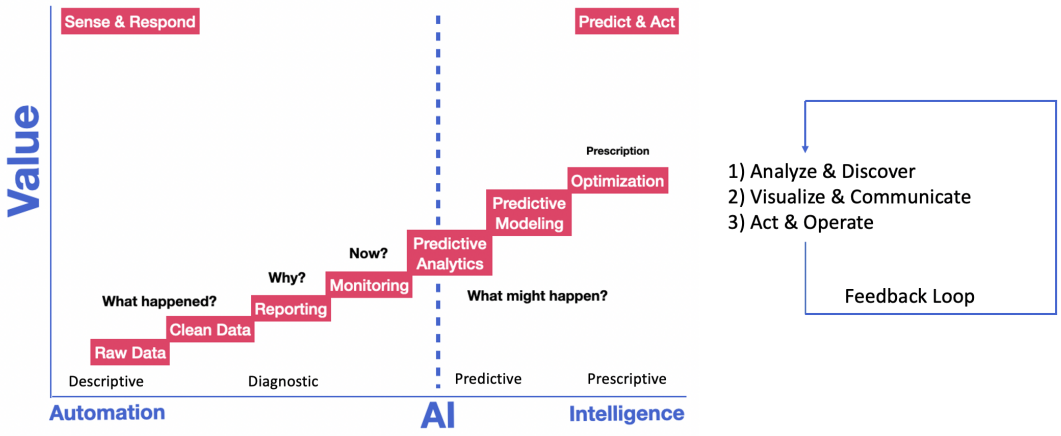
In this graphic, we are using TDA to map an entire complex dataset capturing fraud analytics. On the left is ground truth of where fraud occurs in the network (red) and where it does not exist (blue). Then we take a model developed by a customer and highlight in red where the model predicted fraud to exist. As you can see there isn't a singular problem of false positives. Instead the model has about 10 **TYPES** of false positive each of which must be addressed by the data scientist separately. Doing this without the aid of a graphic like this is very challenging because the underlying math is abstract. But as you can see, we've taken this complex problem and presented a compelling graphic that guides the data scientist to the regions of a complex dataset where their model fails.

Many models are globally optimized and thus, they fail at specific locations in a dataset in different ways. This technique allows the practitioner to balance differences among the dataset for higher accuracy.

7) algorithms have dimensionality barriers or caps which is one reason why we have an activity called dimension reduction --might be worth mentioning this. The point being we use algorithms to access insights among higher dimensionality datasets but those algorithms have limits due to dimensionality. Might even be worth talking about the curse of dimensionality as a transition from low dimensional business intelligence analytics to machine learning for higher dimensionality problems. Here the issue is a human can't manage the rulesets but the machine can. This creates one aspect of the xAI problem.

Hope this is helpful.

Thx,
Brennan



--
Brennan Murphy
Vice President, Defense & National Security
Ayasdi AI

https://www.ayasdi.com/public-sector/