OPINIONS (HTTPS://ANALYTICSINDIAMAG.COM/CATEGORY/ARTICLES/)

8 Explainable AI Frameworks Driving A New Paradigm For Transparency In AI

BY RAM SAGAR (HTTPS://ANALYTICSINDIAMAG.COM/AUTHOR/RAM-SAGAR/)
18/10/2019

Due to the ambiguity in Deep Learning solutions (https://analyticsindiamag.com/8-platforms-you-can-use-to-build-mobile-deep-learning-solutions/), there has been a lot of talk about how to make explainability inclusive of an ML pipeline. Explainable AI refers to methods and techniques in the application of artificial intelligence technology (AI) such that the results of the solution can be understood by human experts. It contrasts with the concept of the "black box" in machine learning and enables transparency.

For example, consider the following two images:

$$
\begin{aligned}
f(\mathbf{x}) \;=\;& 0.293\ \tanh(0.337\,x_1 - 0.329\,x_2 + 0.251\,x_3 - 0.288\,x_4 - 0.297\,x_5 + 0.436\,x_6 + \\
&+\ 0.166\,x_7 - 0.184\,x_8 + 0.219\,x_9 + 0.483\,x_{10} - 0.22\,x_{11} \\
&+\ 0.352\,x_{14} + 0.259\,x_{15} + 0.176\,x_{16} + 0.345\,x_{17} + 0.314\,x_{18} + 0.177\,x_{19} - 0.329\,x_{20} + \\
&-\ 0.363\,x_{21} + 0.216\,x_{22} - 0.148\,x_{23} - 0.043\,x_{24} + 0.316\,x_{25} - 0.068\,x_{26} - 0.421\,x_{27(0)} + \\
&+\ 0.15\,x_{27(1)} - 0.289\,x_{27(2)} - 0.241\,x_{28} + 0.16\,x_{29} + 0.199\,x_{30} - 0.111\,x_{31} - 0.164\,x_{32} + \\
&+\ 0.117\,x_{33} + 0.466\,x_{34} + 0.457\,x_{35} + 0.133\,x_{36} + 0.331\,x_{37} - 0.362\,x_{38} - 0.43\,x_{39} + \\
&-\ 0.491\,x_{40} - 0.155\,x_{41} + 0.371\,x_{42} - 0.05\,x_{43} - 0.177\,x_{44} - 0.044\,x_{45} + 0.225\,x_{46} + \\
&+\ 0.328\,x_{47} - 0.118\,x_{48} - 0.3) + \\
&-\ 1.934\ \tanh(-0.233\,x_0 + 0.174\,x_1 - 0.252\,x_2 - 0.501\,x_3 - 0.125\,x_4 + 0.311\,x_5 - 0.573\,x_6 + \\
&-\ 0.299\,x_7 + 1.123\,x_8 + 0.318\,x_9 - 1.169\,x_{10} + 0.105\,x_{11} - 0.429\,x_{12} - 0.075\,x_{13} + \\
&-\ 0.143\,x_{14} + 0.146\,x_{15} - 0.531\,x_{16} + 0.077\,x_{17} - 0.133\,x_{18} - 0.122\,x_{19} + 0.162\,x_{20} + \\
&-\ 0.08\,x_{21} - 0.496\,x_{22} - 0.21\,x_{23} - 0.113\,x_{24} + 0.485\,x_{25} + 0.575\,x_{26} - 0.126\,x_{27(0)} + \\
&+\ 0.135\,x_{27(1)} + 0.022\,x_{27(2)} - 0.352\,x_{28} - 0.693\,x_{29} + 0.379\,x_{30} + 0.409\,x_{31} - 0.109\,x_{32} + \\
&+\ 0.228\,x_{33} + 0.292\,x_{34} + 0.161\,x_{35} - 0.086\,x_{36} - 0.3\,x_{37} - 0.089\,x_{38} + 0.163\,x_{39} + \\
&-\ 0.074\,x_{40} + 0.31\,x_{41} - 0.849\,x_{42} + 0.14\,x_{43} + 0.754\,x_{44} + 0.291\,x_{45} - 0.533\,x_{46} + 0.273\,x_{47} + \\
&-\ 0.285\,x_{48} - 0.286) + 0.252
\end{aligned}
$$

(https://analyticsindiamag.com/wp-content/uploads/2019/10/Equation-705x382.png)

via Rulex XAI

```
IF (customer_province in {A, B, C, D} AND damage_class in {1} AND Number of days between policy start and date of accident <= 371
   THEN Fraud = Yes
IF (customer_province in {E, B, C, F} AND Customer age > 48 AND Number of days between date of accident and complaint  > 1)
   THEN Fraud = Yes
IF (customer_province in {G, H, I, J, K, L, M, N, B, O, P, Q, R, S})
   THEN Fraud = No
IF (Number of days between date of accident and policy end <=  2)
   THEN Fraud = No
```

(https://analyticsindiamag.com/wp-content/uploads/2019/10/ruledx.png)

via Rulex XAI

The first picture consists of a bunch of mathematical expressions chained together that represent the way inner layers of an algorithm or a neural network function. Whereas the second picture also contains the working of an algorithm but the message is more lucid.

Given an opportunity to choose, any client would prefer the second example. The kind of reputation that ML models have gained over the years has made users skeptical. This problem is more prevalent in use cases where the results are of critical nature. These can be an image recognition algorithm identifying criminals or a model deployed for cancer diagnosis or it can be a recommendation model that pushes certain news and products forward.

The non-transparency of the algorithms can lead to the exploitation of certain groups.

To promote explainable AI, researchers have been developing tools and techniques and here we look at a few which have shown promising results: over the past couple of years:

What-if Tool

TensorFlow team announced the What-If Tool (https://pair-code.github.io/what-if-tool/), an interactive visual interface designed to help visualize datasets and better understand the output of TensorFlow models. To analyze the models deployed. In addition to TensorFlow models, one can also use the What-If Tool for XGBoost and Scikit Learn models.

Once a model has been deployed, its performance can be viewed on a dataset in the What-If tool.

Additionally, one can slice the dataset by features and compare performance across those slices, identifying subsets of data on which the model performs best or worst, which can be very helpful for ML fairness investigations.

LIME

Local Interpretable Model-Agnostic Explanations LIME is an actual method developed by researchers at the University Of Washington to gain greater transparency on what's happening inside an algorithm.

When the number of dimensions is high, maintaining local fidelity for such models becomes increasingly hard. In contrast, LIME solves the much more feasible task of finding a model that approximates the original model locally

LIME (https://arxiv.org/pdf/1602.04938v1.pdf) incorporates interpretability both in the optimization and the notion of interpretable representation, such that domain and task specific interpretability criteria can be accommodated.

LIME, a modular and extensible approach to faithfully explain the predictions of any model in an interpretable manner. The team also introduced SP-LIME, a method to select representative and non-redundant predictions, providing a global view of the model to users

DeepLIFT

DeepLIFT is a method that compares the activation of each neuron to its 'reference activation' and assigns contribution scores according to the difference. It gives separate consideration to positive and negative contributions, DeepLIFT (https://github.com/kundajelab/deeplift) can also reveal dependencies which are missed by other approaches. Scores can be computed efficiently in a single backward pass.

DeepLIFT is on pypi, so it can be installed using pip:

pip install deeplift

Skater

Skater is a unified framework to enable Model Interpretation for all forms of model to help one build an Interpretable machine learning system often needed for real world use-case. Skater is an open source python library designed to demystify the learned structures of a black box model both globally and locally.

Shapley

The Shapley Value SHAP (SHapley Additive exPlanations) is the average marginal contribution of a feature value over all possible coalitions.

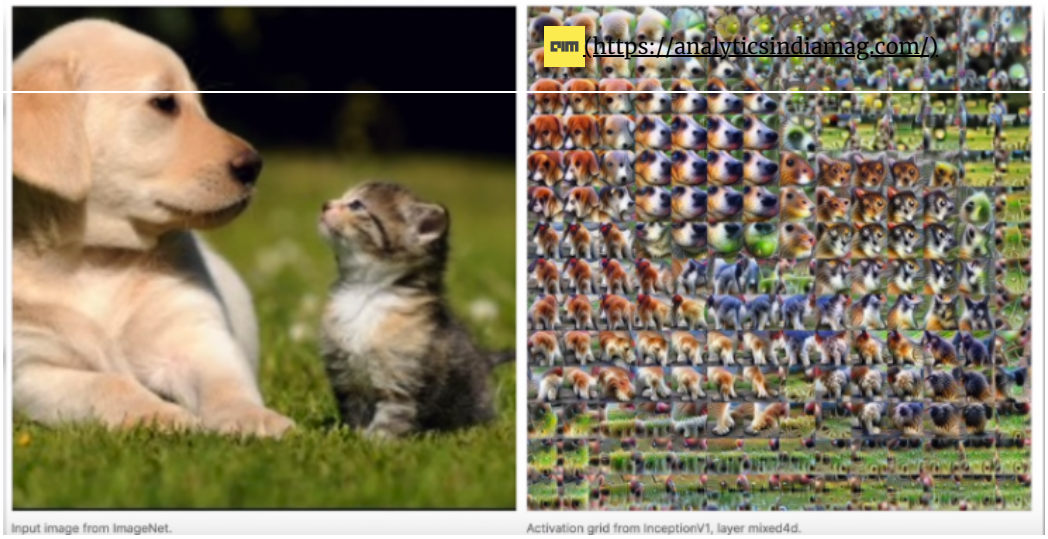
Coalitions are basically combinations of features which are used to estimate the shapley value of a specific feature. It is a unified approach to explain the output of any machine learning model.

SHAP connects game theory with local explanations, uniting several previous methods and representing the only possible consistent and locally accurate additive feature attribution method based on expectations.

AIX360

The AI Explainability 360 (https://github.com/IBM/AIX360) toolkit is an open-source library developed by IBM in support of interpretability and explainability of datasets and machine learning models. The AI Explainability 360 is released as a Python package that includes a comprehensive set of algorithms that cover different dimensions of explanations along with proxy explainability metrics.

Activation Atlases

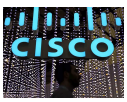


(https://analyticsindiamag.com/wp-content/uploads/2019/10/aas.png)

Google (https://ai.googleblog.com/2019/03/exploring-neural-networks.html) in collaboration with OpenAI (https://blog.openai.com/introducing-activation-atlases/), came up with Activation Atlases, which was a novel technique aimed at visualising how neural networks interact with each other and how they mature with information along with the depth of layers.

This approach was developed to have a look at the inner workings of convolutional vision networks and derive human-interpretable overview of concepts within the hidden layers of a network.

SEE ALSO



(https://analyticsindiamag.com/what-is-the-hiring-process-of-data-scientists-at-cisco/)

OPINIONS (HTTPS://ANALYTICSINDIAMAG.COM/CATEGORY/ARTICLES/)

What Is The Hiring Process Of Data Scientists At Cisco? (https://analyticsindiamag.com/what-is-the-hiring-process-of-data-scientists-at-cisco/)

Rulex Explainable AI

Rulex (https://www.rulex.ai/rulex-explainable-ai-xai/) is a company that creates predictive models in the form of first-order conditional logic rules that can be immediately understood and used by everybody.

Rulex's core machine learning algorithm, the Logic Learning Machine (LLM), works in an entirely different way from conventional AI. The product in designed so that it produces conditional logic rules that predict the best decision choice, in plain language that is immediately clear to process professionals. Rulex rules make every prediction fully self-explanatory.

And unlike decision trees and other algorithms that produce rules, Rulex rules are stateless and overlapping.

The Need For Transparency In Models

Scroll To Top

(https://analyticsindiamag.com/wp-content/uploads/2019/10/fouad-ghazizadeh-jRHU8rNXrlA-unsplash.jpg)

Data quality and its accessibility are two main challenges one will come across in the initial stages of building a machine learning pipeline.

But there can be problems associated with the information that is deployed into the model such as:

- an incorrect model gets pushed
- incoming data is corrupted
- incoming data changes and no longer resembles datasets used during training.

Today, developers are eve using deep learning to analyze another deep learning network, allowing them to understand the inner working of a model and the kind of influence as it flows to layers above or below and ultimately back down to the underlying data source.

Operating in the dark has made AI less trustworthy as a solution. Providing a solution to any problem is the end of the story to any ML model. However, for the practitioners it is crucial to explain their results in the most intuitive to their clients.

The European Union's General Data Protection Regulation (GDPR), which went into effect in 2018, insists on having high level data protection for consumers and harmonizes data security regulations within the European Union (EU).

Today, the companies have to inform subjects about any personal data collection and processing and obtain their consent before collecting such data as the GDPR threatens the use of traditional machine learning AI technology for automated decisions.

According to article 14 of GDPR, when a company uses automated decision-making tools, it must provide meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.

For a safer, more reliable inclusion of AI, a seamless blend of human and artificial intelligence is needed. Human intervention should also be considered in developing techniques that allow practitioners to easily evaluate the quality of the decision rules in use and reduce false positives.

What Do You Think?

Add a comment...

Facebook Comments Plugin

Subscribe to our Newsletter

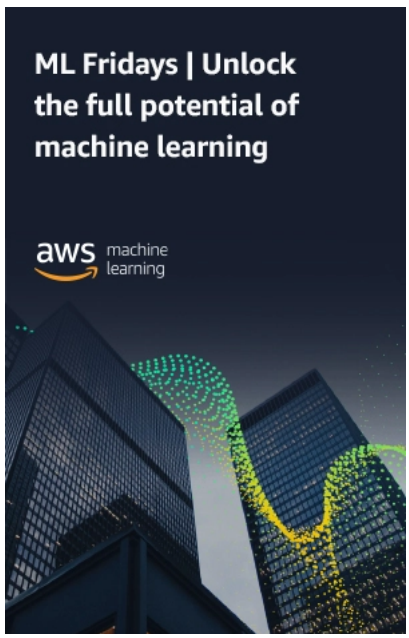
Get the latest updates and relevant offers by sharing your email.

ENTER YOUR EMAIL **SUBSCRIBE NOW**

Join Our Telegram Group. Be part of an engaging online community. Join Here (https://t.me/joinchat/NJLxnhZB7GkX3CPvjs9QGQ).

RAM SAGAR (HTTPS://ANALYTICSINDIAMAG.COM/AUTHOR/RAM-SAGAR/)

I have a master's degree in Robotics and I write about machine learning advancements.
email:ram.sagar@analyticsindiamag.com



Scroll To Top

OUR UPCOMING EVENTS

SKILLUP 2021 | Data Science Education Fair | 22-23rd April | Register here>> (https://skillup.analyticsindiasummit.com/)

Rising 2021 (https://rising.analyticsindiasummit.com/) | Women in AI Conference | May 21 & 22 | Virtual

RELATED POSTS

OPINIONS (HTTPS://ANALYTICSINDIAMAG.COM/CATEGORY/ARTICLES/)

The Struggle For Data Privacy (https://analyticsindiamag.com/the-struggle-for-data-privacy/)

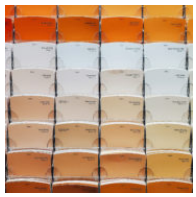
11/03/2021 · 3 MINS READ

(https://analytics
struggle-for-
data-privacy/)

DEVELOPERS CORNER (HTTPS://ANALYTICSINDIAMAG.COM/CATEGORY/DEVELOPERS_CORNER/)

Hands-on Guide to Interpret Machine Learning with SHAP (https://analyticsindiamag.com/hands-on-guide-to-interpret-machine-learning-with-shap/)

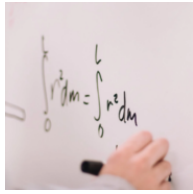
06/03/2021 · 6 MINS READ



(https://analytics
on-guide-to-
interpret-
machine-
learning-with-
shap/)

DEVELOPERS CORNER (HTTPS://ANALYTICSINDIAMAG.COM/CATEGORY/DEVELOPERS_CORNER/)

Guide To AI Explainability 360: An Open Source Toolkit By IBM (https://analyticsindiamag.com/guide-



(https://analytics
to-ai-

Scroll To Top

to-ai-explainability-360-an-open-source-toolkit-by-ibm/) (https://analyticsindiamag.com/)

29/01/2021 · 16 MINS READ

(https://analytics
explainability-360-an-open-source-toolkit-by-ibm/)

DEVELOPERS CORNER (HTTPS://ANALYTICSINDIAMAG.COM/CATEGORY/DEVELOPERS_CORNER/)

Hands-On Guide To Adversarial Robustness Toolbox (ART): Protect Your Neural Networks Against Hacking (https://analyticsindiamag.com/adversarial-robustness-toolbox-art/)

07/01/2021 · 7 MINS READ

(https://analytics
robustness-toolbox-art/)

PEOPLE (HTTPS://ANALYTICSINDIAMAG.COM/CATEGORY/INTERVIEWS/)

Model Explainability Validates AI For Safety-Critical Systems, Says Prashant Rao, MathWorks India (https://analyticsindiamag.com/model-explainability-validates-ai-for-safety-critical-systems-says-prashant-rao-mathworks-india/)

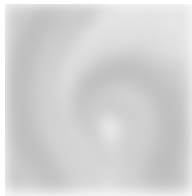
29/12/2020 · 5 MINS READ

(https://analytics
explainability-validates-ai-for-safety-critical-systems-says-prashant-rao-mathworks-india/)

OPINIONS (HTTPS://ANALYTICSINDIAMAG.COM/CATEGORY/ARTICLES/)

A Greater Foundation For The Triumph Of Deep Learning With XAI (https://analyticsindiamag.com/a-greater-foundation-for-the-triumph-of-deep-learning-with-xai/)

21/12/2020 · 5 MINS READ



(https://analytics
greater-foundation-for-the-triumph-of-deep-learning-with-xai/)

Scroll To Top