6 Explainable AI (XAI) Frameworks for Transparency in AI

#machinelearning #ai #datascience

amananandrai Jan 14 · *Updated on Feb 25* · 5 min read

Artificial Intelligence (AI) is affecting our daily lives in many ways. Facial Recognition Systems, Artificial Assistants, Predictive Models are used nearly everywhere. AI is finding its use in many industries such as Education, Healthcare, Automobile, Manufacturing, and Law Enforcement. The decisions and predictions being made by AI-enabled systems are becoming much more important and in many cases, critical to life, and death. This is especially true for AI systems used in healthcare, driverless cars, or even drones being used during the war.

The working of these AI models is however not understandable by the common man as well as professionals. If an Artificial Neural Network makes a prediction from an image whether it is of a cat or a dog it is not obvious that based on what characteristic or features this decision is made. This decision however is not as critical to our life as predicting whether an image of a tumor cell is that of a malignant or benign tumor. We would obviously like to know whether based on what parameters this decision is made.

In healthcare, the explainability of AI is of utmost importance. Previously, Machine Learning and Deep Learning models were treated as black boxes that took some input and made decision to give some output but based on what parameters these decisions are made is not obvious. Now with the increasing use of AI in our daily life and AI making decisions affecting our life and death in cases like autonomous cars, cancer prediction softwares the need for **Explainability** in AI has increased.

As humans, we must be able to fully understand how decisions are being made so that we can trust the decisions of AI systems. The lack of explainability and trust hampers our ability to fully trust AI systems. We want computer systems to work as expected and produce transparent explanations and reasons for decisions they make. This is known as **Explainable AI (XAI)**.

♡ 6 5 8 •••

Explainable AI is a new and budding field in the area of AI and Machine Learning. It is very important to build trust among humans about the decisions made by AI models. It is only possible by making the black box of ML models more transparent. Explainable AI frameworks are tools which generate reports about how the model works and tries to explain their working. In this blog, we will be discussing about 6 explainable AI frameworks

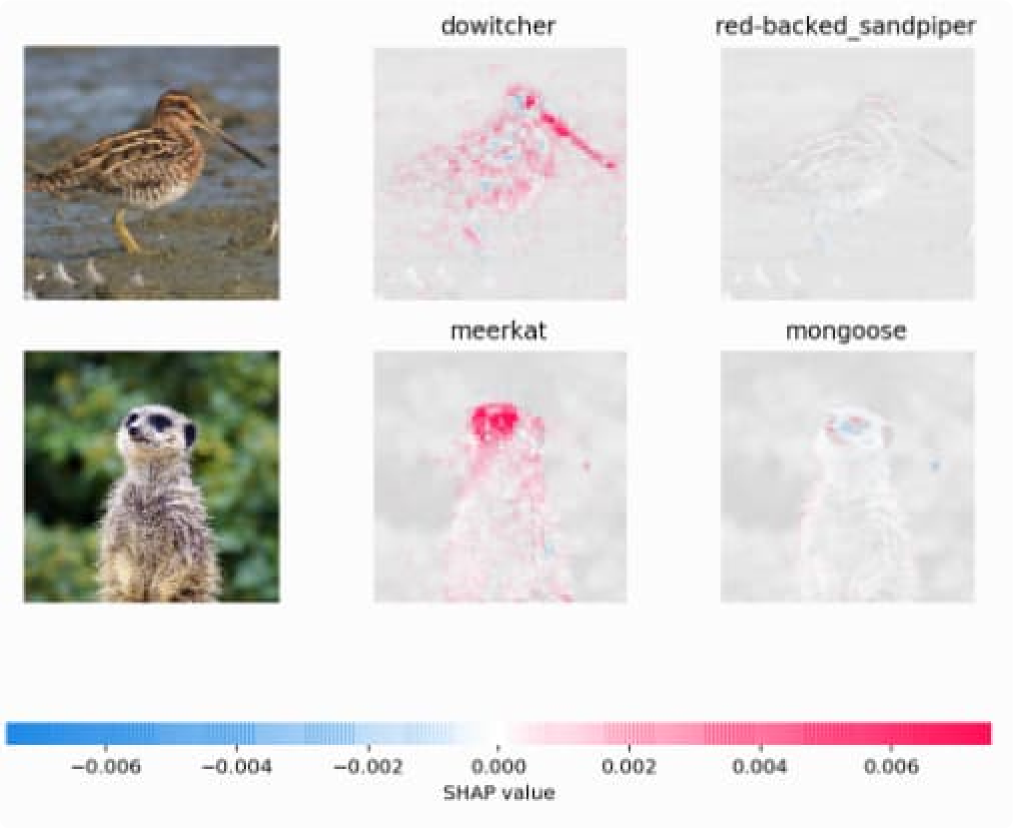
- SHAP
- LIME
- ELI5
- What-if Tool
- AIX360
- Skater

1- SHAP

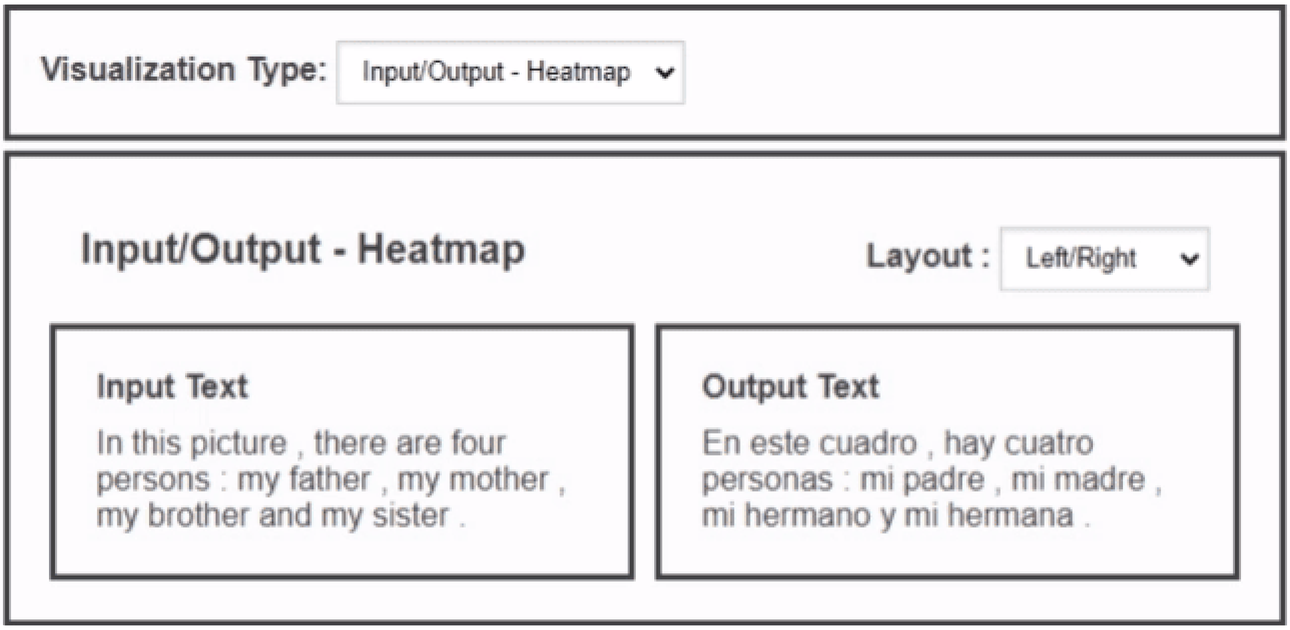
SHAP stands for **SH**apley **A**dditive ex**P**lanations. It can be used for explaining various types of models like simple machine learning algorithms like linear regression, logistic regression, tree-based models, and also more complex models like deep learning models for image classification and image captioning and various NLP tasks like sentiment analysis, translation, and text summarization. It is a model agnostic method to explain the models based on game theory's shapley values. It explains how the different features affect the output or what contribution do they have in the outcome of the model.

An example of a sentiment analysis explainer using SHAP is given here.

Some other examples are shown below



Explanation of an image classification model



Visualization Type: Input/Output - Heatmap

Input/Output - Heatmap

Layout : Left/Right

Input Text

In this picture , there are four persons : my father , my mother , my brother and my sister .

Output Text

En este cuadro , hay cuatro personas : mi padre , mi madre , mi hermano y mi hermana .

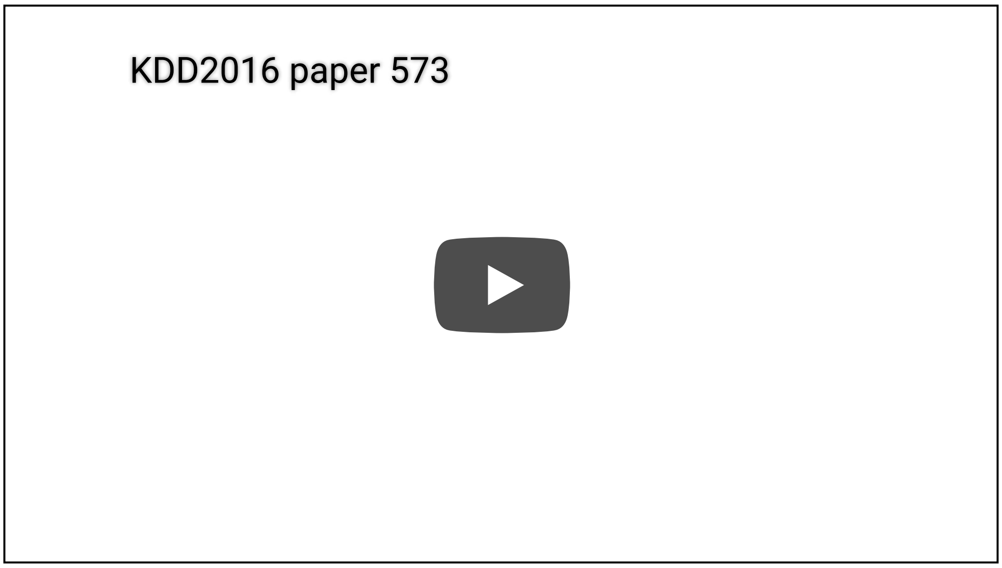
Example of text translation explanation

To know more visit the link- https://shap.readthedocs.io/

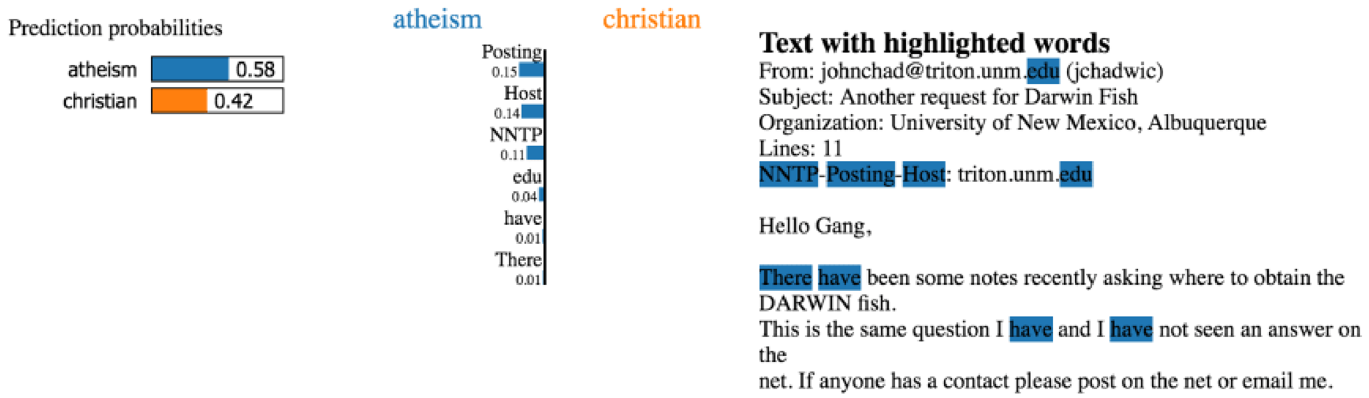
but is faster in terms of computation. The output of LIME is a list of explanations,

reflecting the contribution of each feature to the prediction of a data sample. Lime is able to explain any black box classifier, with two or more classes. All it requires is that the classifier implements a function that takes in raw text or a numpy array and outputs a probability for each class. Support for scikit-learn classifiers is built-in.

A video about LIME



Some of the screenshots showing the explanations given by LIME



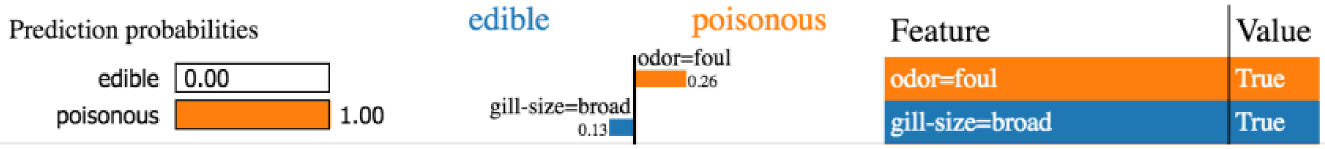
Text with highlighted words

From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu

Hello Gang,

There have been some notes recently asking where to obtain the
DARWIN fish.
This is the same question I have and I have not seen an answer on
the
net. If anyone has a contact please post on the net or email me.

For two class textual data



Feature	Value
odor=foul	True
gill-size=broad	True

Tabular data

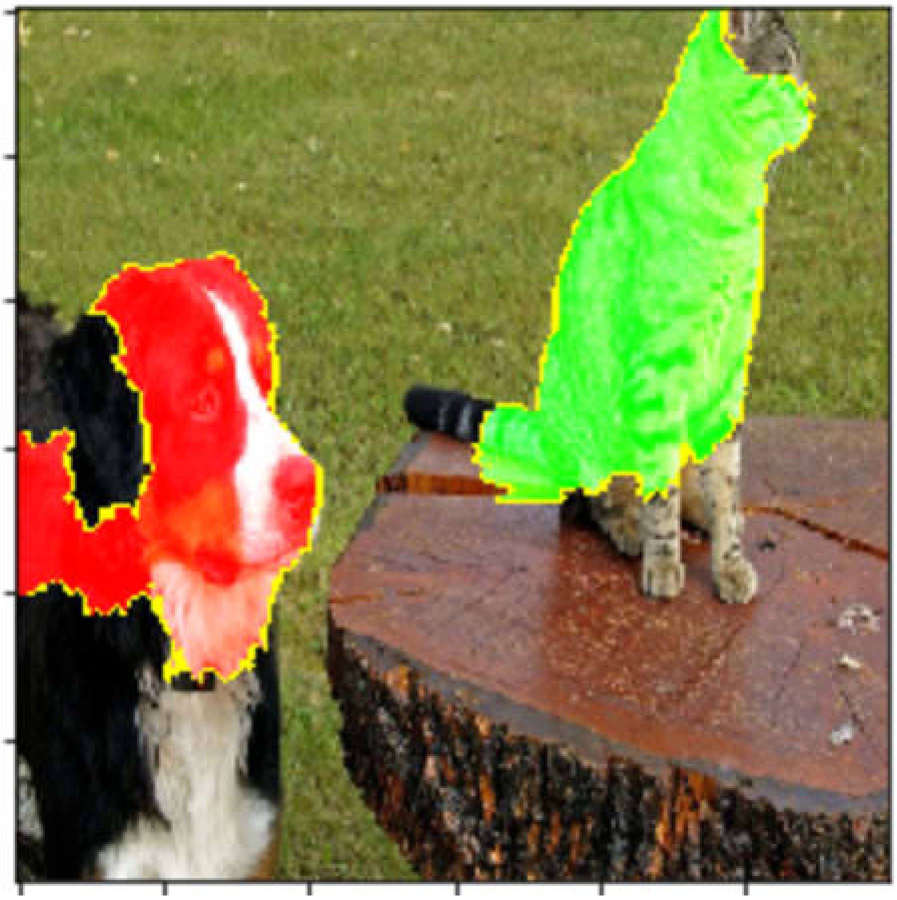


Image explaining cat prediction green(+ve) and red(-ve)

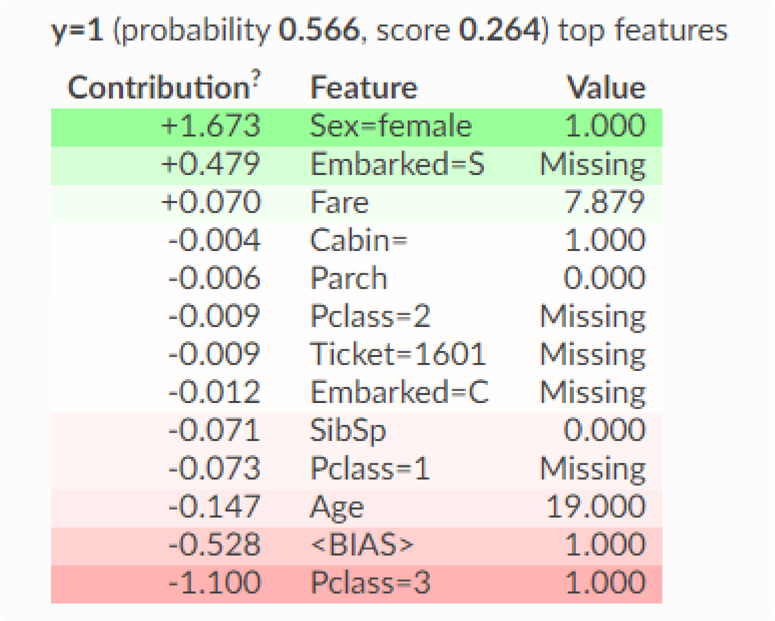
Link- https://github.com/marcotcr/lime

3- ELI5

ELI5 is a Python package which helps to debug machine learning classifiers and explain their predictions. It has support for many ML frameworks like scikit-learn, Keras, XGBoost, LightGBM, and CatBoost.

There are two main ways to look at a classification or a regression model:

1) inspect model parameters and try to figure out how the model works globally;
2) inspect an individual prediction of a model, try to figure out why the model makes the decision it makes.



y=1 (probability **0.566**, score **0.264**) top features

Contribution?	Feature	Value
+1.673	Sex=female	1.000
+0.479	Embarked=S	Missing
+0.070	Fare	7.879
-0.004	Cabin=	1.000
-0.006	Parch	0.000
-0.009	Pclass=2	Missing
-0.009	Ticket=1601	Missing
-0.012	Embarked=C	Missing
-0.071	SibSp	0.000
-0.073	Pclass=1	Missing
-0.147	Age	19.000
-0.528	<BIAS>	1.000
-1.100	Pclass=3	1.000

Example showing the importance of various features in Titanic Dataset

Link- https://eli5.readthedocs.io/

4- What-if Tool

Whatif Tool (WIT) is developed by Google to understand the working of ML trained models. Using WIT, you can test performance in hypothetical situations, analyze the importance of different data features, and visualize model behavior across multiple models and subsets of input data, and for different ML fairness metrics. The What-If Tool is available as an extension in Jupyter, Colaboratory, and Cloud AI Platform notebooks. It can be used for different tasks like binary classification, multi-class classification and regression. It can be used with various types of data like Tabular, Image and Text data. It can be used along with SHAP and LIME. It can also be used with Tensor Board.

Link- https://pair-code.github.io/what-if-tool/

5- AIX360

AIX360 or AI Explainability 360 is an extensible open source toolkit that can help you

Link- https://aix360.mybluemix.net/

6- 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-cases. It is an open source python library designed to demystify the learned structures of a black box model both globally(inference on the basis of a complete data set) and locally(inference about an individual prediction).

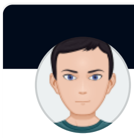
Link- https://github.com/oracle/Skater

Discussion (0)

Subscribe

DEV Add to the discussion

Code of Conduct • Report abuse



amananandrai

Data Science and Machine Learning Enthusiast

Follow

LOCATION
Ballia, U.P., India IN

EDUCATION
NIT Bhopal (M. Tech.)

JOINED
Mar 25, 2020

More from amananandrai

♡ 6   5   8   •••

#machinelearning #nlp

What are your favourite Datascience and Machine Learning blogs?

#datascience #machinelearning #watercooler #discuss

List of Evaluation Metrics for Classification and Regression

#machinelearning

♡ 6 5 8 •••