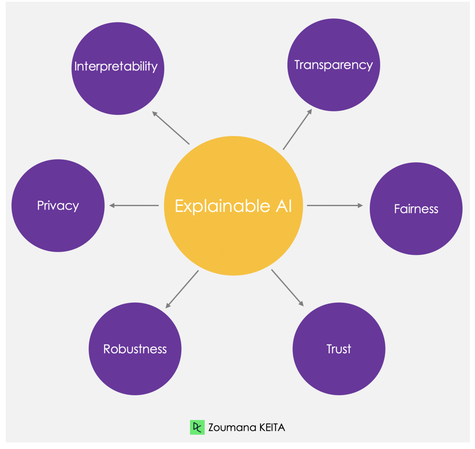
Explainability for Artificial Intelligence on Healthcare

Sannu Kumar Nayak

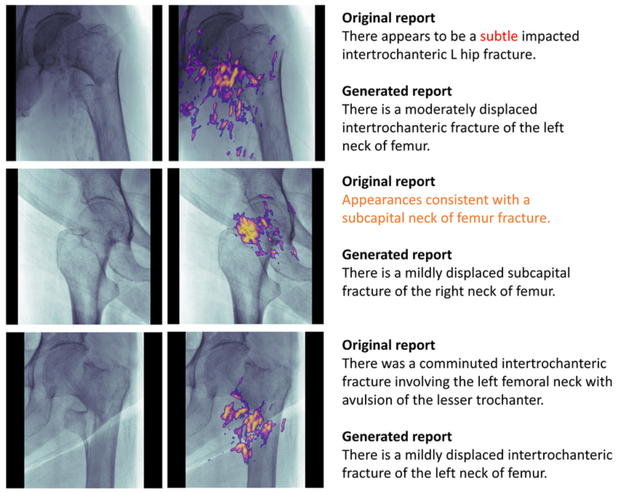
Cs460 Machine Learning, SCoS, NISER

Explainability

Explainable artificial intelligence (XAI) is a powerful tool in answering critical How? and Why? questions about AI systems and can be used to address rising ethical and legal concerns.



Explainable AI refers to the set of processes and methods that allows human users to comprehend and trust the results and output created by machine learning algorithms.



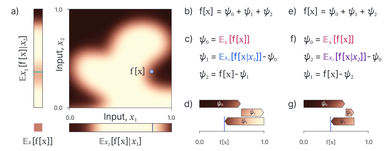
In the healthcare domain, for instance, researchers have identified explainability as a requirement for AI clinical decision support systems because the ability to interpret system outputs facilitates shared decision-making between medical professionals and patients and provides much-needed system transparency.

During the 1970s to 1990s, symbolic reasoning systems, such as MYCIN could represent, reason about, and explain their reasoning for diagnostic, instructional, or machine-learning (explanation-based learning) purposes. MYCIN, developed in the early 1970s as a research prototype for diagnosing bacteremia infections of the bloodstream, could explain which of its hand-coded rules contributed to a diagnosis in a specific case.Similarly, GUIDON added tutorial rules to supplement MYCIN's domain-level rules so it could explain the strategy for medical diagnosis.

Explaining SHapley Additive exPlanations (SHAP)

SHAP uses the game theory concept of Shapley values to optimally assign feature importances.

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



Shapley additive explanations describe the model output f[$\mathbf{x}_i$] for a particular input $\mathbf{x}_i$ as an additive sum:

$$\mathrm{f}[\mathrm{x}_i] = \psi_0 + \sum_{d=1}^{D} \psi_d(1)$$

of D contributing factors ψ_D associated with the D dimensions of the input. In other words, the change in performance from a baseline ψ_0 is attributed to a sum of changes ψ_d associated with the input dimensions.

Consider the case where there are only two input variables, choosing a particular ordering of the input variables and a constructing the explanation piece by piece. So, we might set:

$$\begin{aligned}
\psi_0 &= \mathbb{E}_{\mathbf{x}}\left[f[\mathbf{x}]\right] \\
\psi_1 &= \mathbb{E}_{\mathbf{x}}\left[f[\mathbf{x}]|x_1\right] \left(\mathbb{E}_{\mathbf{x}}\left[f[\mathbf{x}]\right]\right) \\
\psi_2 &= \mathbb{E}_{\mathbf{x}}\left[f[\mathbf{x}]|x_1, x_2\right] - \left(\mathbb{E}_{\mathbf{x}}\left[f[\mathbf{x}]\right] - \mathbb{E}_{\mathbf{x}}\left[f[\mathbf{x}]|x_1\right]\right). \qquad (2)
\end{aligned}$$

Shapley Additive Explanations

The idea of Shapley additive explanations is to compute the values ψ_d by taking a weighted average of the ψ_i over all possible orderings. If the set of indices is given by $\mathcal{D} = \{1, 2, \ldots, D\}$, then the final Shapley values are

$$\psi_d[f[\mathbf{x}]] = \sum_{\mathcal{S}\subseteq\mathcal{D}} \frac{|\mathcal{S}|!(D - |\mathcal{S}| - 1)!}{D!} \left(\mathbb{E}\left[f[\mathbf{x}]|\mathcal{S}\right] \mathbb{E}\left[f[\mathbf{x}]|\mathcal{S}_{\setminus d}\right]\right). \qquad (5)$$

This computation takes every subset of variables that contains x_d and computes the expected value of the function given this subset takes the particular values for that data point with and without x_d itself. This result is weighted and contributes to the final value ψ_d . The particular weighting (i.e., the first term after the sum) can be proven to be the only one that satisfies the properties of (i) local accuracy (the Shapley values sum to the true function output) (ii) missingness (an absent feature has a Shapley value/ attribution of zero, and (iii) consistency (if the marginal contribution of a feature increases or stays the same, then the Shapley value should increase or stay the same).

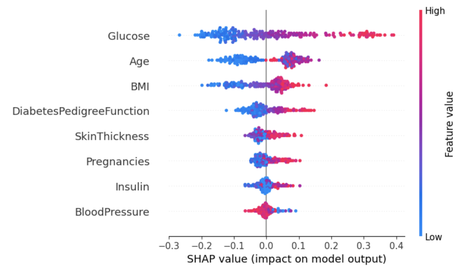
A Sample Code

```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split

import shap
explainer = shap.TreeExplainer(model)
shap_values = explainer.shap_values(X)
shap.summary_plot(shap_values, X_test)
```

Applicaation of SHAP

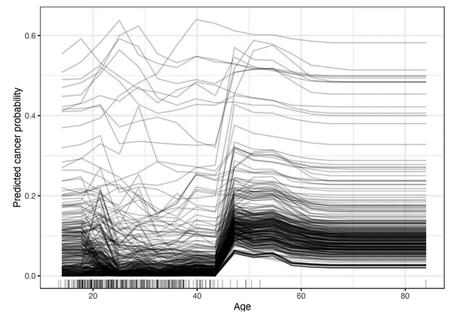
We will explore the SHAP interpretation method using the famous "Pima Indians Diabetes Database" to predict whether a patient has diabetes or not.



From the above graphic:

Y-axis represents the features ranked by their average absolute SHAP values.
X-axis represents SHAP values. Positive values for a given feature push the model's prediction closer to the label being examined (label=1). In contrast, negative values push towards the opposite class (label=0).
An individual with a high glucose (red dots) level is likely to be diagnosed with diabetes (positive outcome), while a low glucose level leads to not being diagnosed with diabetes. Similarly, aging patients are more likely to be diagnosed with diabetes. However, the model seems uncertain about the diagnosis for younger patients.

Conclusion



CE plot of cervical cancer probability by age. Each line represents one person. For most people. there is an increase in predicted cancer probability with increasing age. For some women with a predicted cancer probability above 0.4, the prediction does not change much at higher age.

References

[1] S Prince.
Explainability i: local post-hoc explanations, 2022.
Accessed: 2024-Apr-2.

[2] Violet Turri.
What is explainable ai?
Carnegie Mellon University, Software Engineering Institute's Insights (blog), Jan 2022.
Accessed: 2024-Apr-2.

sannukumar.nayak@niser.ac.in