Model Explainability in AI

ARITRA GANGULY

05. 07. 2022

Machine learning and deep learning are becoming ubiquitous due to:

- The ability to solve complex problems in a variety of different domains.
- The growth in the performance and efficiency of modern computing resources.
- The widespread availability of large amounts of data.

However, as the size and complexity of problems continue to increase, so does the complexity of the machine learning algorithms applied to these problems. The inherent and growing complexity of machine learning algorithms limits the ability to understand what the model has learned or why a given prediction was made, acting as a barrier to the adoption of machine learning. Additionally, there may be legal or regulatory requirements to be able to explain the outcome of a prediction from a machine learning model, resulting in the use of biased models at the cost of accuracy.

Machine learning explainability (MLX) is the process of explaining and interpreting machine learning and deep learning models.

MLX can help machine learning developers to:

- Better understand and interpret the model's behavior.
  - Which features does the model consider important?
  - What is the relationship between the feature values and the target predictions?
- Debug and improve the quality of the model.
  - Did the model learn something unexpected?
  - Does the model generalize, or did it learn something specific to the training dataset?
- Increase trust in the model and confidence in deploying the model.

MLX can help users of machine learning algorithms to:

- Understand why the model made a certain prediction.
  - Why was my bank loan denied?

Some useful terms for MLX:

- **Explainability:** The ability to explain the reasons behind a machine learning model's prediction.
- **Interpretability:** The level at which a human can understand the explanation.
- **Global Explanations:** Understand the general behavior of a machine learning model as a whole.
- **Local Explanations:** Understand why the machine learning model made a specific prediction.
- **WhatIf Explanations:** Understand how changes in the value of features affect the model's prediction.
- **Model-Agnostic Explanations:** Explanations treat the machine learning model and feature pre-processing as a black box instead of using properties from the model to guide the explanation.

Ways to interpret a Model.

There are two ways to interpret the model - Global vs. Local interpretation.

Global Interpretation	**Local Interpretation**
It helps in understanding how a model makes decisions for the overall structure.	It helps in understanding how the model makes decisions for a single instance.
Using global interpretation, we can explain the complete behavior of the model.	Using local interpretation, we can explain the individual predictions.
Global interpretation help in understanding the suitability of the model for deployment.	Local interpretation helps in understanding the behavior of the model in the local neighborhood.
Example - Predicting the risk of disease in patients.	Example - Understanding why a specific person has a high risk of disease.
• SHAP (SHapley Additive exPlanations) [Documentation] [GitHub] • LIME (Local Interpretable Model-Agnostic Explanations) [Documentation] [GitHub] • ELI5 [Documentation] [GitHub]	• PDP (Partial Dependency Plot) • ICE (Individual Conditional Expectation)

References:

- Interpretable Machine Learning - Christoph Molnar
- Model Explainability - Oracle Blog
- An End-to-End Guide to Model Explainability