Vemburoj Konar
28
D16AD

AAI
=
Exp-9
=

DATE:

Aim: Implement and analyze the working of Local Interpretable Model agnostic Explainations (LIME) supervised model.

Theory:

LIME is a methodology developed to provide human-interpretable explainations for the predictions of complex machine learning models, regardless of the underlying algorithm used.

1) Local Interpretability
LIME focuses on providing explaination at the local level meaning explainations are tailored to individual predictions rather than the entire dataset.

2) Perturbation-Based Explaination
It generates explainations by perturbing the input features around the instance of interest and observing the resulting changes in the model's prediction.

3) Weighting and Importance
LIME assigns weights to the perturbed samples based on their proximity to the instance are given higher weights, indicating their greater influence on the local explaination.

DATE:

4) Interpretable Explanation

Once the interpretable model is trained, LIME provides explanations in the form of feature contributions or decision rules that explain how each input feature influences the model's prediction for the instance of interest

Advantages:

1) Model Agnostic

It can be applied to any ML models regardless of underlying architecture

2) Local Explanations

By focusing on local explanations, LIME provides insights into how the model behaves for specific instances, allowing for better understanding and validation of individual predictions

3) Interpretability

LIME generates explanations that are interpretable to humans', enabling stakeholders to trust and validate model predictions, identify potential biases, and gain insights into the decision making process.

Conclusion:

LIME offers a principled approach to explaining the predictions of complex Machine learning models in a human-interpretable manner. LIME serves as a valuable tool for improving the accountability, fairness, and reliability of machine learning systems in various domains.