Explainable AI:

What is our model learning? What input feature is important?

AI models transition from black box (not interpretable even by experts) to glass box (interpretable ML).

Builds trust by providing explanation for predictions (important in case of healthcare)

InterpretML Package:

Open-source Python package that contains different interpretability algorithms.

Glassbox Model: linear regression, logistic regression, decision trees, simple models, trained as part of the package, supports both local and global explanations.

Blackbox Model: svm,random forest,neural networks, complicated models, trained outside of the package, supports only local explanation.

Tools: 1. LIME 2. SHAP

*(Note: EBM/Explainable Boosting Machine is a glassbox model that can give high accuracy similar to random forest)*

XAI Categorization:

Based on Agnosticity

Model Agnostics -> It can be applied to any ML model.

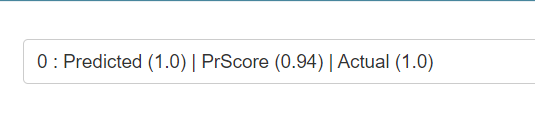
Model Specific -> Designed for specific ML model.

Based on Scope

Global explanation -> For the whole model.

Local explanation -> For individual predictions.

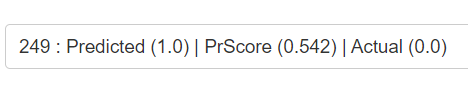
Glassbox Model:

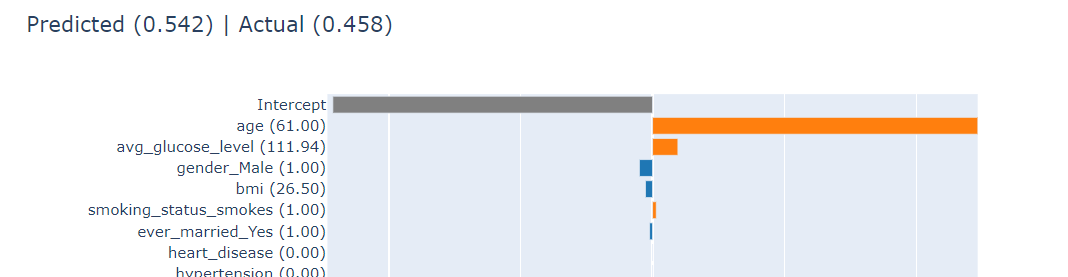




Age, history of heart disease, private work etc. is positively impacting the prediction of stroke. A good bmi is negatively impacting the prediction of stroke.

PrScore (Probability Score/Predicted Probability) --> The model has predicted the label to be 1 with a probability of 0.94. The actual target is 1. Thus, the model has mode the correct prediction.





PrScore (Probability Score/Predicted Probability) --> The model has predicted the label to be 1 with a probability of 0.542. The actual target is 0(that had a predicted probability of 0.458). Thus, the model has mode the incorrect prediction.

Methods/Tools for explaining Black Box Algorithm:

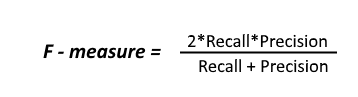
Lime (Local Interpretable Agnostic-Model Explanation)

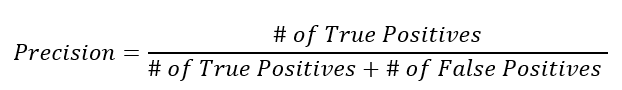
It is not model specific.

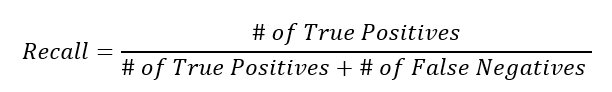
Valid explanation is derived by zooming in (local interpretation). Aims to explain by local approximation (converts an overall complex model to a simple localized model).

Additional:

F1 score:







Important when the data has class imbalance( I.e category A has frequency 99 while category B has frequency 1).

Reference:

[InterpretML: Another Way to Explain Your Model | by Noga Gershon Barak | Towards Data Science](https://towardsdatascience.com/interpretml-another-way-to-explain-your-model-b7faf0a384f8)

[Welcome to The Much Anticipated Interpret Documentation!](https://interpret.ml/docs/intro.html)

[I](https://interpret.ml/docs/getting-started.html)[interpret-ml - Explain Machine Learning Models And Their Predictions [Python] (coderzcolumn.com)](https://coderzcolumn.com/tutorials/machine-learning/interpret-ml-explain-machine-learning-models-and-their-predictions)

[xai-series/01\_interpretable\_models.py at master · deepfindr/xai-series (github.com)](https://github.com/deepfindr/xai-series/blob/master/01_interpretable_models.py)