

Explainability Approaches

By Timing

Direct
Interpretability

Decision Trees
Linear Models
Rule Lists

Post-hoc
Explanation

LIME
SHAP
Anchors

By Scope

Local
(Per Instance)

Feature Attribution
Counterfactuals
Influence Functions

Global
(Model-Wide)

Model Distillation
Probing Classifiers
TCAV

By Model Access

Model-
Specific

Attention Viz
LRP
DeepLIFT

Model-
Agnostic

LIME
SHAP
Anchors

