

SHAP Values for Model Interpretation

SHapley Additive exPlanations - unified framework for interpretability

Shapley Theory

- From game theory (Nobel Prize)
- Fair contribution of each feature
- Considers all feature combinations
- Unique solution with nice properties

SHAP Algorithms

- TreeSHAP (fast for trees)
- DeepSHAP (for neural nets)
- KernelSHAP (model-agnostic)
- LinearSHAP (for linear models)

Visualization Types

Waterfall Plots

Single prediction

Summary Plots

Global importance

Dependence Plots

Feature effects

Interaction Plots

Feature interactions