MLDL

SAND2020-7793C

Machine Learning and Deep Learning Conference 2020

Assessing the Accuracy of ML Explanations for Model Credibility

- Michael R. Smith/5952
- Erin Acquesta/5954, Arlo Ames/6331, Rich Field/5953, Trevor Maxfield/9302, Blake Moss/9312, Megan Nyre-Yu/6671, Ahmad Rushdi/1463, Charles Smutz/9312, Mallory Stites/6672

- LDRD

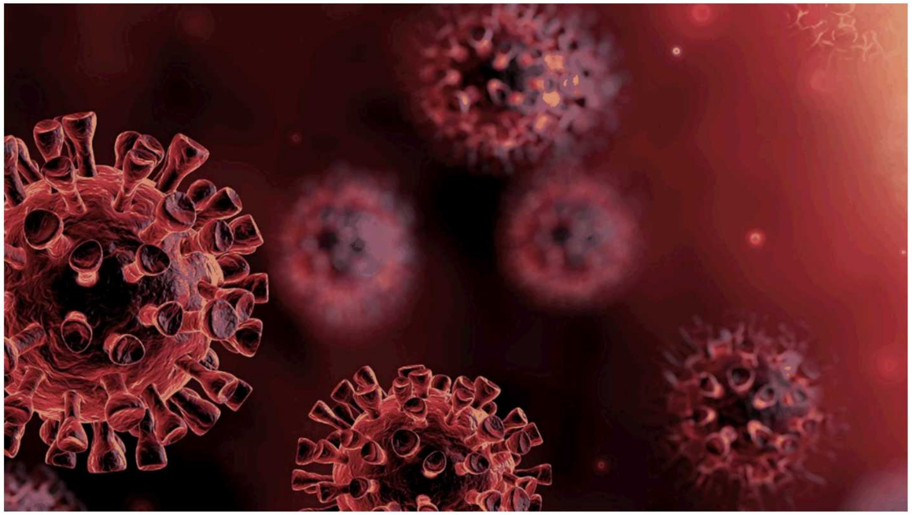
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Outline

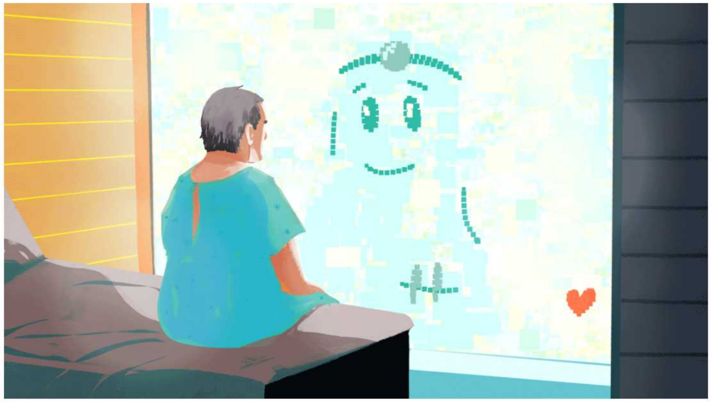
- The Need for Model Credibility
  - The Impetus on National Security Labs
  - The need for Credible Explainability
- Current Explainability methods
- Sensitivity Analysis Guided Explainability
- Preliminary Results

Impetus for National Security Labs

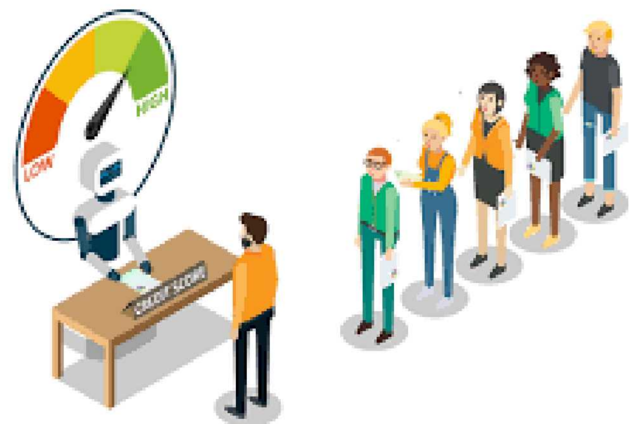
- Model high consequence applications with uncertainty for decision makers to ingest
- Several established mathematical fields for rigorous analysis:
  - Quantifying uncertainty bounds
  - Sensitivity of the model to input parameters
  - Etc.
- Policies and regulations may be the result of the analyses



The Need for Credible Explainability



- ML is being used in an increasingly number of high-consequence applications.
  - ML explainability has emerged as field that seeks to build trust.
    - Computational shortcuts
    - Assumes some understanding of machine learning
    - Lack verifiable foundations



- **Can we trust the explanation?**
  - Explainability is unique to ML
  - Other fields provide rigor to model validation and credibility that is lacking in ML explainability

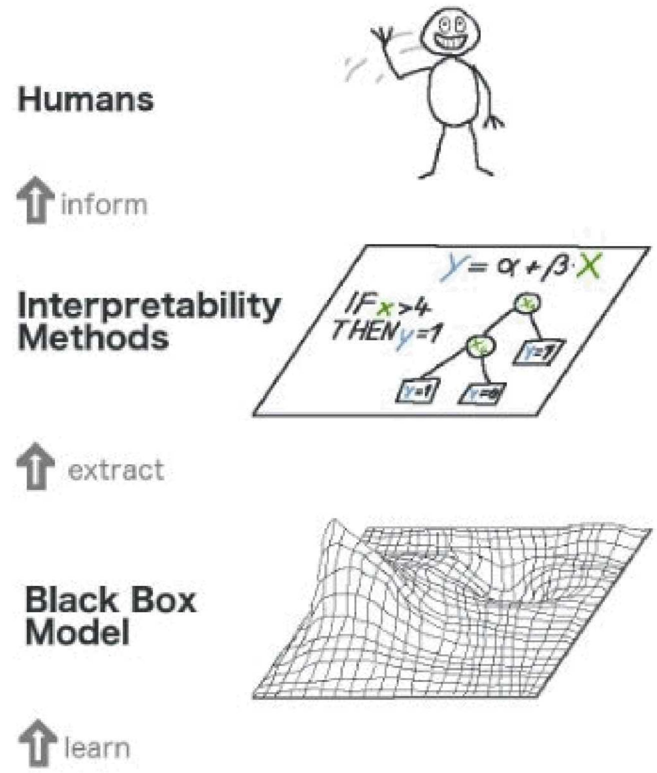
Outline

- The Need for Model Credibility
- Current Explainability methods
  - Overview of Explainability
  - LIME
  - SHAP
  - LIME and SHAP deficiencies
- Sensitivity Analysis Guided Explainability
- Preliminary Results

Current Explainability Methods

Explainability: describe the decision process that a model considers for a prediction

- Global VS Local
- Interpretable models
  - Can inspect the model
  - Models that are relatively easy to interpret (linear regression models, shallow decision trees)
- White-box/Integrated
  - Can inspect the model, but the model is sufficiently complex
  - Gini-importance for decision trees
  - Gradient-based methods for deep learning models (Saliency maps)
- **Black-box/Post-hoc**
  - **Do not inspect the model**
  - **Often perturb the data and observe how the output changes**
  - **Create a surrogate model that is interpretable or provides and explanation**

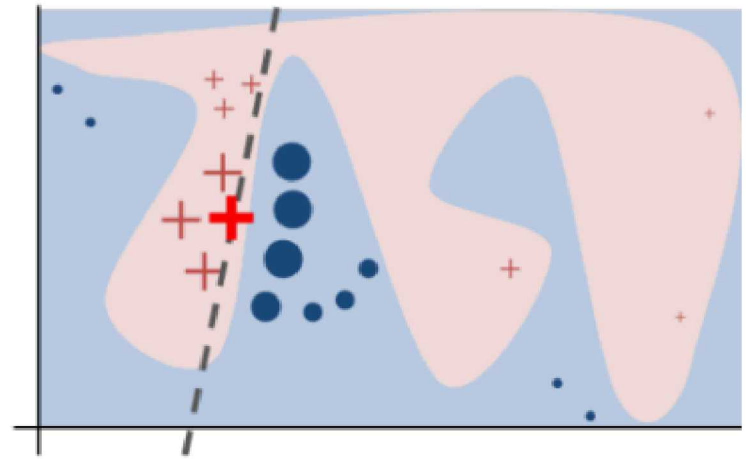


Black-Box Explanation Methods

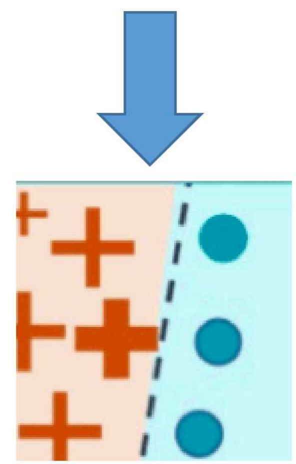
Local Interpretable Model-agnostic Explanations (LIME)

- Perturbation-based

1. Sample data from a Gaussian and rescale
2. Calculate the distance between the sampled instances and the instance being explained
3. Make predictions; record output
4. Fit a weighted (step 2) linear model on this data set
    1. LIME explanation = regression coefficients * feature values



Complex Non-linear



Simple Linear

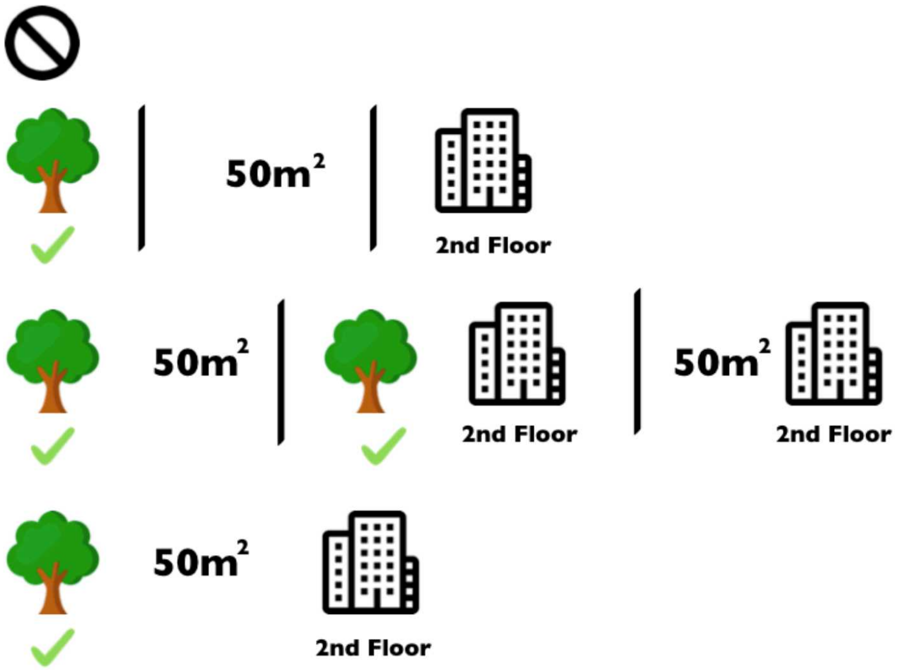
Black-Box Explanation Methods

SHapley Additive exPlanations (SHAP)

- Based on cooperative game theory

1. For each *feature*:
   1. For each instance:
      1. Replace the feature value with a randomly selected value
      2. Make prediction on the modified instance
      3. Calculate the distance between average prediction and the modified instance
2. SHAP value = Average difference from the average prediction

$$g(x') = \phi_0 + \sum_{j=1}^{M} \phi_j$$



Black-Box Explanation Methods

Dependencies and Assumptions

- Dependent on a process for sampling the data
- Require distance on output—how much it changes
- Assumes independence and linearity

Observed Deficiencies

- **Descriptive Accuracy**: Match when features are removed
- **Instability**: Produces different explanations on the same input
- **Completeness**: Generate explanations for all possible input vectors
- **Efficiency**: Can be slow to calculate especially as the dimensionality increases

No agreed upon definition of explainability or what constitutes an explanation

Outline

- The Need for Model Credibility
- Current Explainability methods
- Sensitivity Analysis Guided Explainability
  - SAGE overview
  - Global Sensitivity Analysis
  - Experimental Design for Sensitivity Analysis
  - Experimental Design for ML Explainability
- Preliminary Results

Do Current Methods Provide Credibility?

- **V&V:** Provides credibility in comp sim models
- **Goal of explainability**: provide credibility by describing the decision process that a model considers for a prediction
- **Problem**: How to provide model credibility for a machine learned model?
- **Approach**: Use existing mathematical frameworks to evaluate explanations
- **SAGE LDRD:** Sensitivity Analysis Guided Explainability
  - Use sensitivity analysis (SA) techniques to understand how inputs affect a model's output
    - Are current SA implementations sufficient for ML data types?
  - Use the results from the SA to explain a model's decision making process
  - Evaluate the impact of the explanations on Enterprise Security

Global Sensitivity Analysis

The <u>apportionment</u> for the contributions of input uncertainties on output uncertainty.

Modeling Flow Diagram

Nominal Parameter Values $\theta = \{\alpha, \beta, \gamma\}$ → Model of the Real World $f(x, \theta)$ → Deterministic Outcome $y \in \mathbb{R}^v$

Example: Infectious Disease Modeling

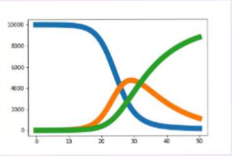
α = seroprevalence
β = infection rate
γ = [infectious duration]$^{-1}$

→

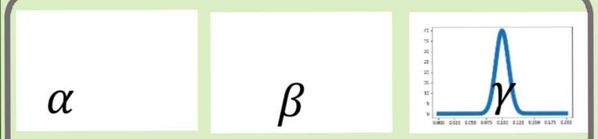
$$x = \begin{bmatrix} S \\ I \\ R \end{bmatrix} \qquad \frac{\mathrm{d}x}{dt} = f(x, \theta), \quad x_0 = g(\theta)$$

→

$y = F(x)$

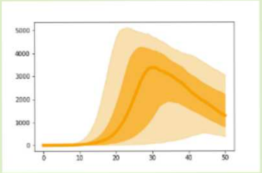


Uncertainty Quantification of Model Forecast for the Infection Rate Curve



→

$$\begin{bmatrix} [\alpha_1, \beta_1, \gamma_1] \\ [\alpha_2, \beta_2, \gamma_2] \\ \cdots \\ [\alpha_s, \beta_s, \gamma_s] \end{bmatrix} \rightarrow f(x, \theta) \rightarrow \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_s \end{bmatrix} \rightarrow$$



Sensitivity of the Infection Rate Curve with respect to the Parameters of the Model

$$S_\alpha = \frac{Var(\alpha)}{Var(y)} \qquad S_{\alpha\beta} = \frac{Cov(\alpha, \beta)}{Var(y)}$$

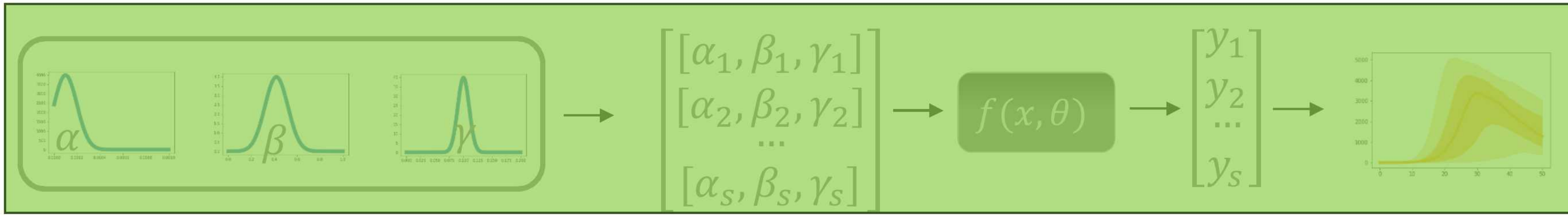
Total Variance of y

$S_{\beta\gamma}$
S_β
$S_{\beta\alpha}$
S_α

Sum of ALL other combined contributions

Experimental Design for Sensitivity Analysis

Experimental Design is a scientific approach for identifying the inputs to a *process* that are most influential to the outcome of that process; following particular design decisions.



Inputs:
Uncertainty in Parameters

Process:
Mathematical Model

Outcome:
Uncertainty in Model Output

Design Decision I

Sampling

Sampling sufficient discrete realizations that preserve the statistics and introduces only marginal standard error.

Design Decision II

Controlled/Uncontrolled Random Behavior

Controlled: Sources of Variance
Uncontrolled: Random behaviors inherent to the model

Design Decision III

Quantity of Interest (QoI)

For the intended use case, what output from the model maps to quantitative metric for that intended purpose.

$$S_\alpha = \frac{Var(\alpha)}{Var(y)} \qquad S_{\alpha\beta} = \frac{Cov(\alpha, \beta)}{Var(y)}$$

Experimental Design for ML Explainability

Original Modeling Flow Diagram

Nominal Parameter Values $\theta = \{\alpha, \beta, \gamma\}$ → Model of the Real World $f(x, \theta)$ → Deterministic Outcome $y \in \mathbb{R}^v$

Translation to ML Classification

Training Data $\mathcal{T} = [X|Y]$ $X \in \mathbb{R}^{m \times n}, Y \in \mathbb{R}^{m \times 1}$

↓

New (unlabeled) Observations $x \in \mathbb{R}^{1 \times n}$ → Machine Learned Model $f(x, \theta(\mathcal{T}))$ → Class Prediction $y \in \{A, B, C\}$

Inputs: Uncertainty in Features	**Process**: Machine Learned Model	**Outcome**: Uncertain Model Predictions
Sampling	**Controlled/Uncontrolled Random Behavior**	**Quantity of Interest (QoI)**
Preserving the statistical properties of the training data: non-Gaussian, discrete, correlated, and sparse	Running sufficient replicates for the random behavior of stochastic machine learned models.	What is the appropriate QoI for which a sensitivity analysis will provide insight for ML explainability?

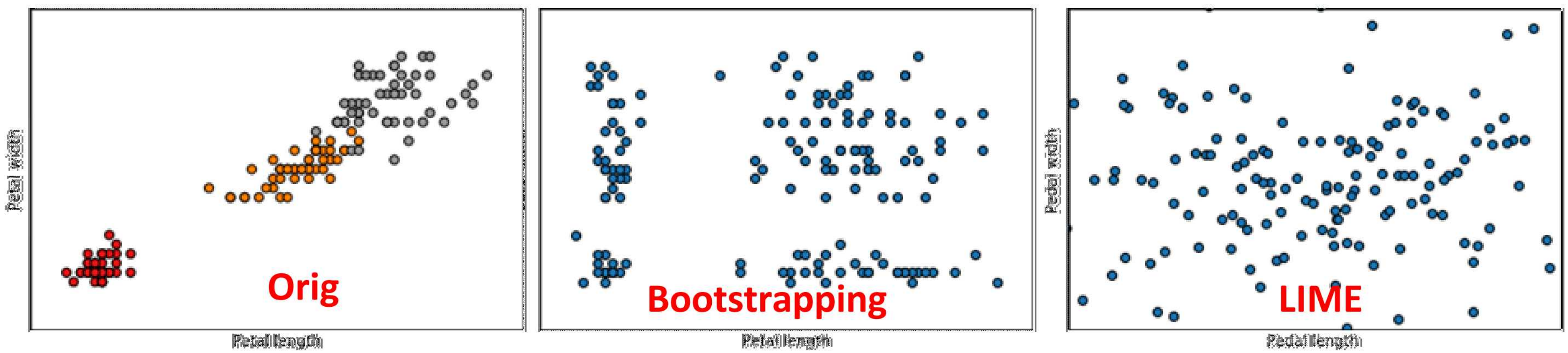
Methods to apportion the influence of sources of input uncertainty across output uncertainty, accounting for higher-order interactions in a model and input correlations.

Outline

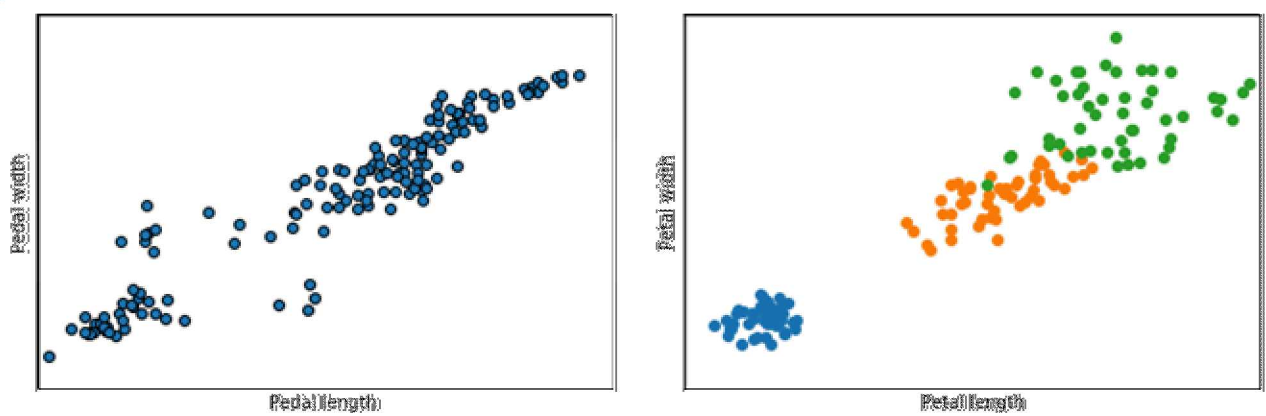
- The Need for Model Credibility
- Current Explainability methods
- Sensitivity Analysis Guided Explainability
- Preliminary Results
    - Correlation Preserving Sampling
    - Quantity of Interest
    - Correlation Does Matter

Correlation Preserving Sampling

1. Current sampling approaches can create unrealistic data points (Do not preserve correlations)
   1. Features 3 and 4 from the iris data set (correlated features)



2. Developed Sampling methods that preserve correlations and generate realistic data points

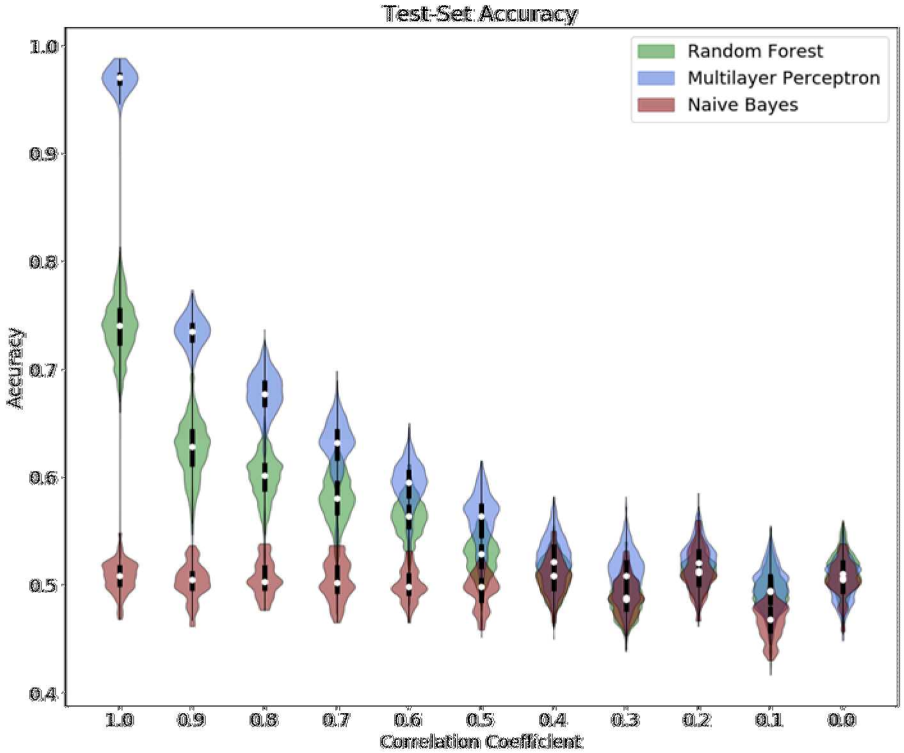


Quantity of Interest

1. Core problem: using variance from a categorical variable
2. For models with high support, the confidence function will be high vast majority of time.
3. Our most promising approaches are:
   1. Model confidence (or similar) recognizing that SA only applies when there is some amount of classifier confusion (poorly supported samples)
   2. Introspection (examining branch purity, NN weights, etc.)
   3. Distance from prototypes approach. Shortcoming of this approach is that metric isn't strictly based on output.
   4. Use ensemble/surrogate model based approach

Correlation Does Matter

1. Scientific approach to the debunking the myth that correlated variables only provide redundant information.
    1. Used synthetic data to control the amount of correlation as the distinguishing characteristic between classes
    2. Naïve Bayes is the baseline for linear relationships
2. Most explainability methods assume independence
    1. Incongruent explanations for the learned model
    2. LIME uses a linear model
    3. SHAP makes independence and linear assumptions
    4. Tested with quadradic regression
3. Still an open research question in SA



SAGE Explanations

1. Current explainability methods lack rigor in verifying their correctness and have several known deficiencies
2. SAGE seeks to use established mathematical frameworks to improve the validity of explainability methods
3. Several open research gaps in SA and applying SA to ML:
    1. How to sample the data that 1) is realistic and 2) cause output variation?
        1. The data is a combination Continuous, Discrete, and Categorial
        2. How to define variance for categorical variables?
    2. How to measure the output variance?
        1. Variance for a categorical variable
    3. How to apportion input variance to output variance preserving correlations and higher-order interactions?

Thank You!!

We Look forward to, and encourage, continued engagements!

E-mail: wg-sage-ldrd@sandia.gov

BACKUP SLIDES

Black-Box Explanation Methods

Local Interpretable Model-agnostic Explanations (LIME)

- Perturbation-based

1. Sample data from a Gaussian and rescale
2. Calculate the distance between the sampled instances and the instance being explained
3. Make predictions; record output
4. Fit a weighted (step 2) linear model on this data set
    1. LIME explanation = regression coefficients * feature values

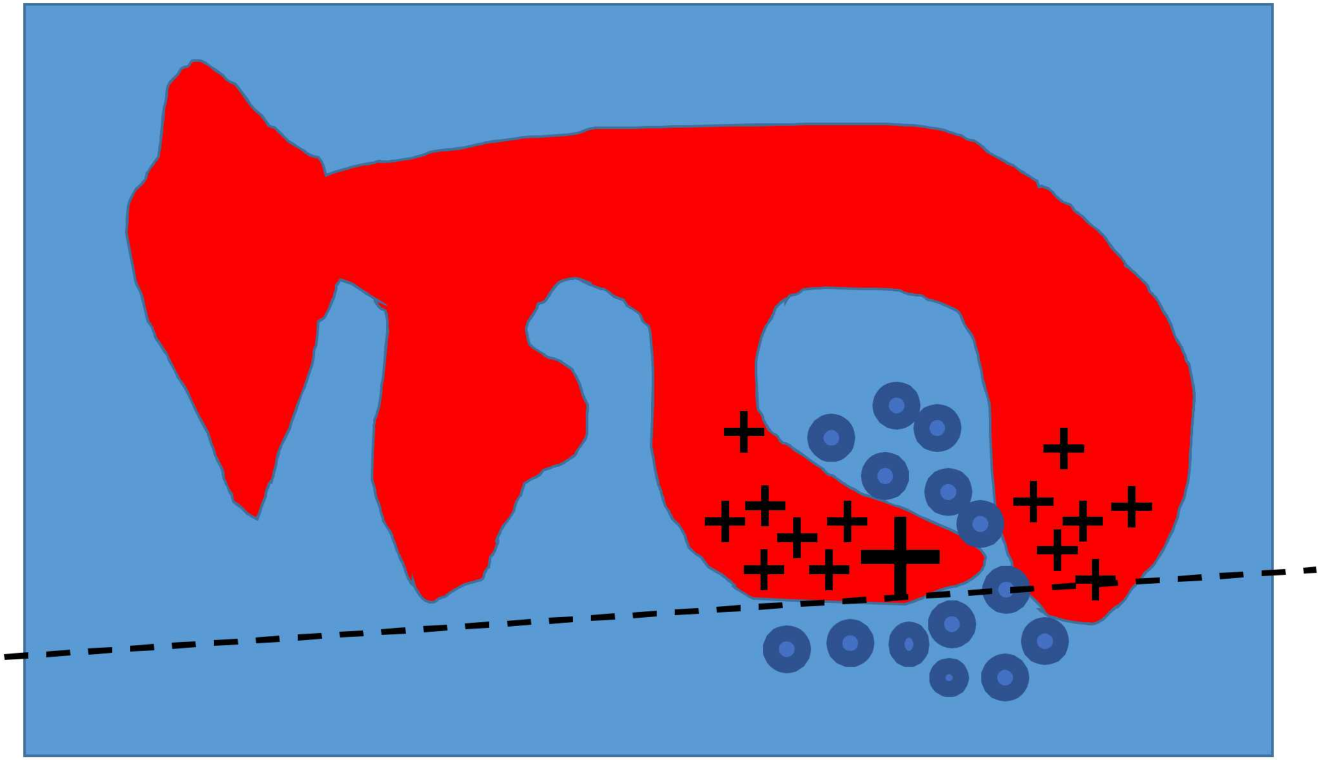
SHapley Additive exPlanations (SHAP)

- Based on cooperative game theory

1. For each *feature*:
    1. For each instance:
        1. Replace the feature value with a randomly selected value
        2. Make prediction on the modified instance
        3. Calculate the distance between average prediction and the modified instance
2. SHAP value = Ave difference from the average prediction

$$g(x') = \phi_0 + \sum_{j=1}^{M} \phi_j$$

Current Explainability Methods



Current Explainability Methods