INL/MIS-24-77221-Revision-0

Improving Machine Learning Explainability with a Graphical User Interface

March 2024

Cody McBroom Walker

INL/MIS-24-77221-Revision-0

Improving Machine Learning Explainability with a Graphical User Interface

Cody McBroom Walker

March 2024

Idaho National Laboratory
Idaho Falls, Idaho 83415

http://www.inl.gov

Prepared for the
U.S. Department of Energy
Under DOE Idaho Operations Office
Contract DE-AC07-05ID14517

Cody Walker, PhD

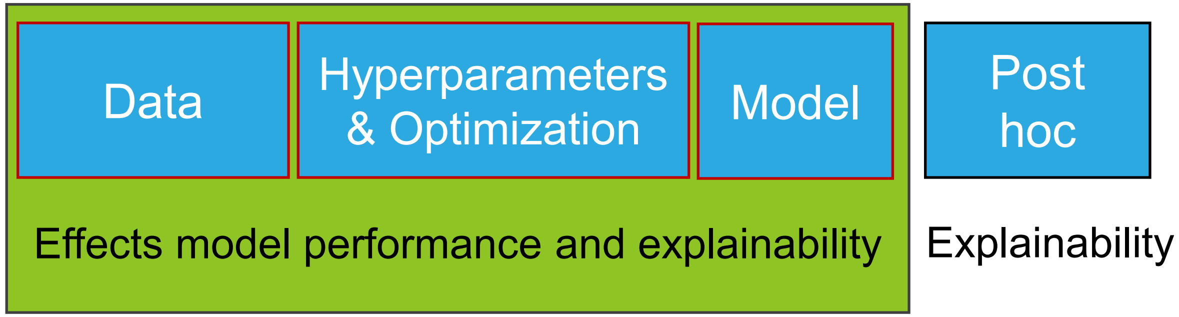
March 28, 2024

Improving Machine Learning Explainability with a Graphical User Interface

Idaho National Laboratory

Explainability comes in different forms depending on which model you use.

- How do we explain these models to the user?
- How much would you have to explain to go from an input to an output?



Data

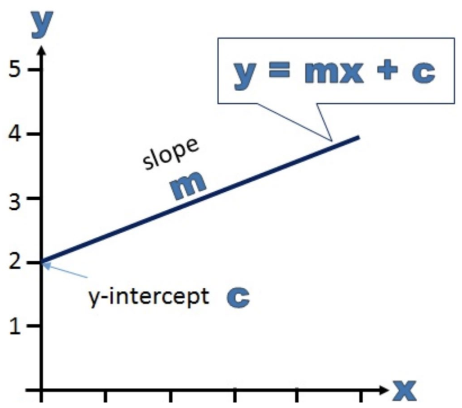
Hyperparameters & Optimization

Model

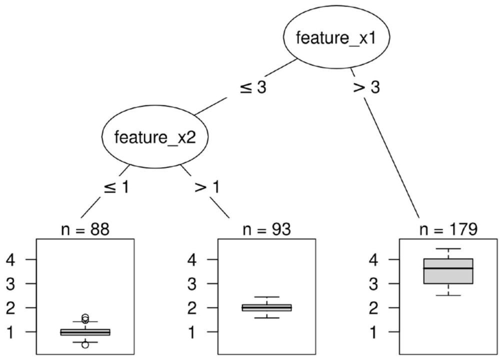
Post hoc

Effects model performance and explainability

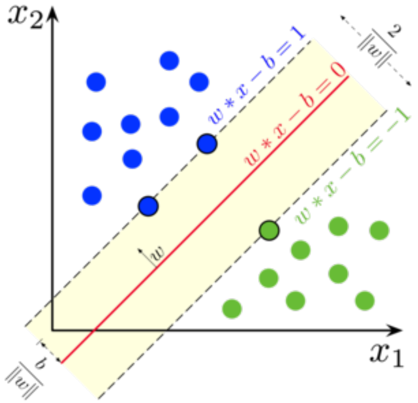
Explainability



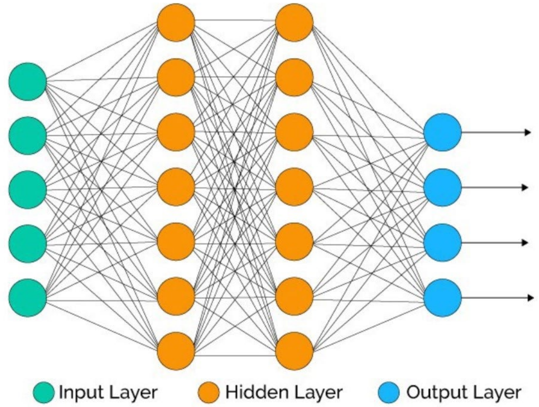
2 [m, c]



2 [2 splits]



Support vectors, kernel trick and hyperplane.



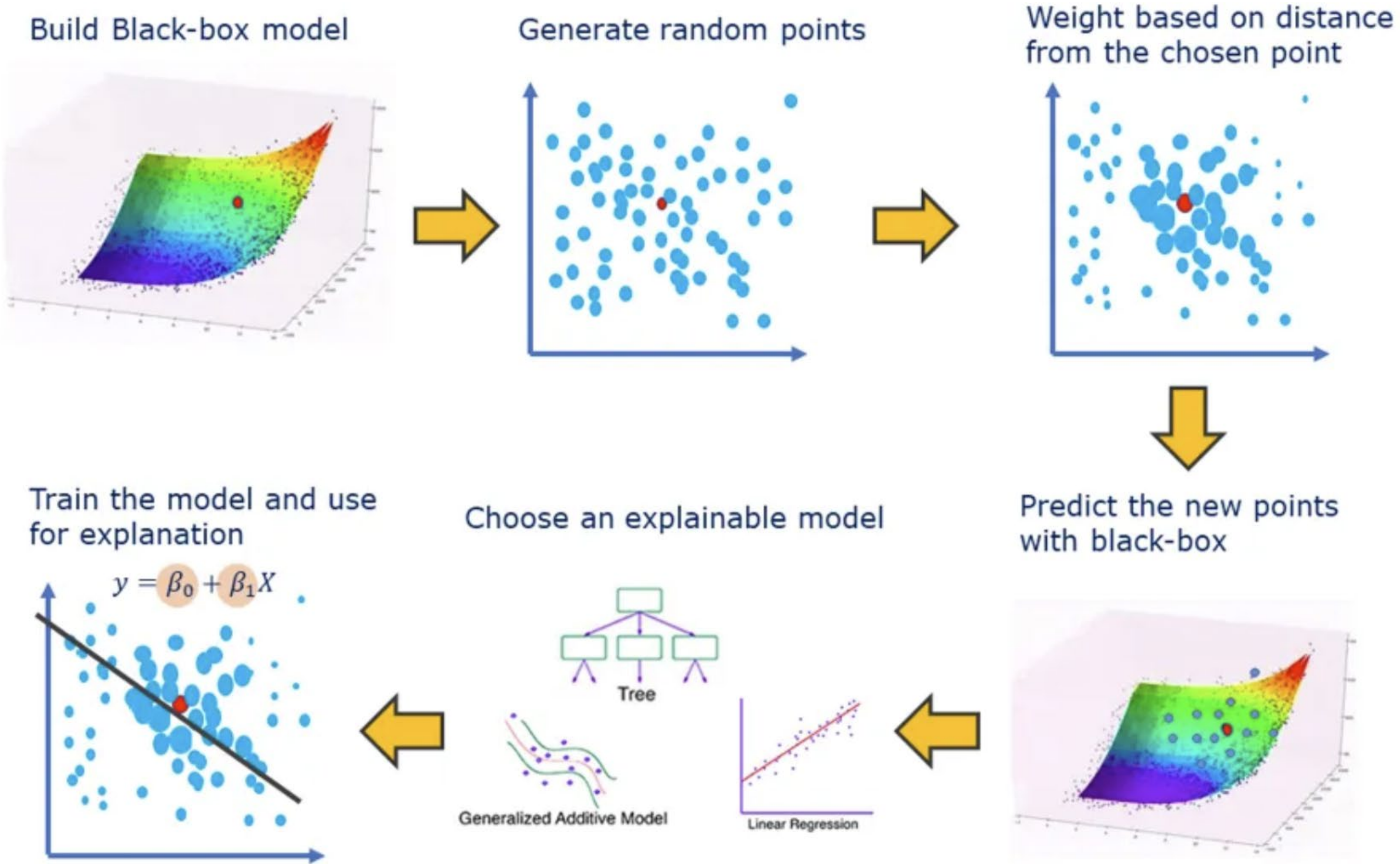
Number of weights, biases, & connections.

LIME is a post-hoc method for black-box models.

- Local interpretable model-agnostic explanations (LIME) can be used for any model.

- LIME is only valid **locally**.

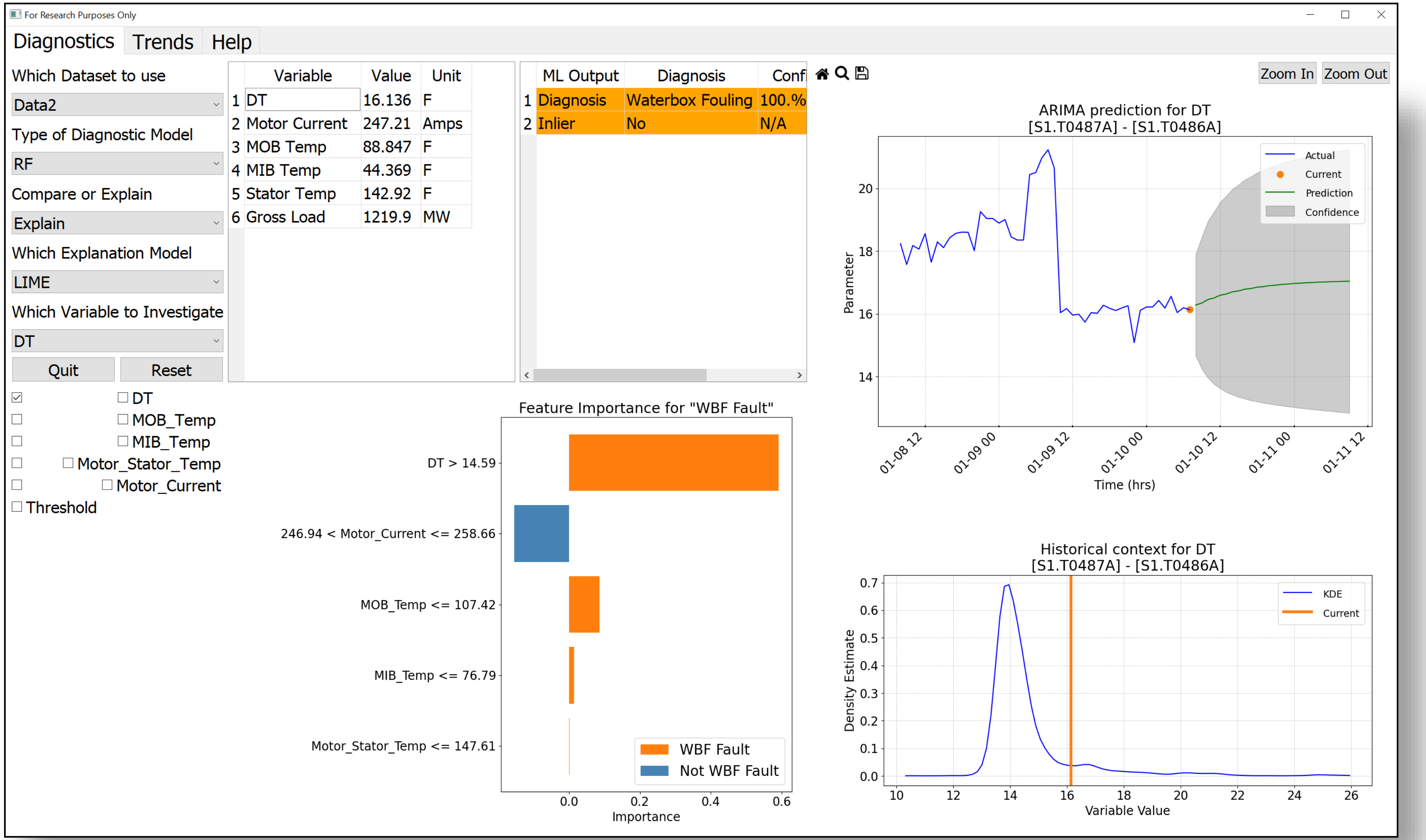
- SHAP (Shapley Additive Explanations) are another common post-hoc method used to increase explainability.



Build Black-box model
Generate random points
Weight based on distance from the chosen point
Predict the new points with black-box
Choose an explainable model
Tree
Generalized Additive Model
Linear Regression
Train the model and use for explanation
$y = \beta_0 + \beta_1 X$

Giorgio Visani, 2020 "LIME: explain Machine Learning predictions."Accessed 2024. https://towardsdatascience.com/lime-explain-machine-learning-predictions-af8f18189bfe

Model confidence, prognostics, explainability, and historical context all provide evidence for the conclusion.



Which Dataset to use: Data2

Type of Diagnostic Model: RF

Compare or Explain: Explain

Which Explanation Model: LIME

Which Variable to Investigate: DT

☑ □ DT

□ □ MOB_Temp

□ □ MIB_Temp

□ □ Motor_Stator_Temp

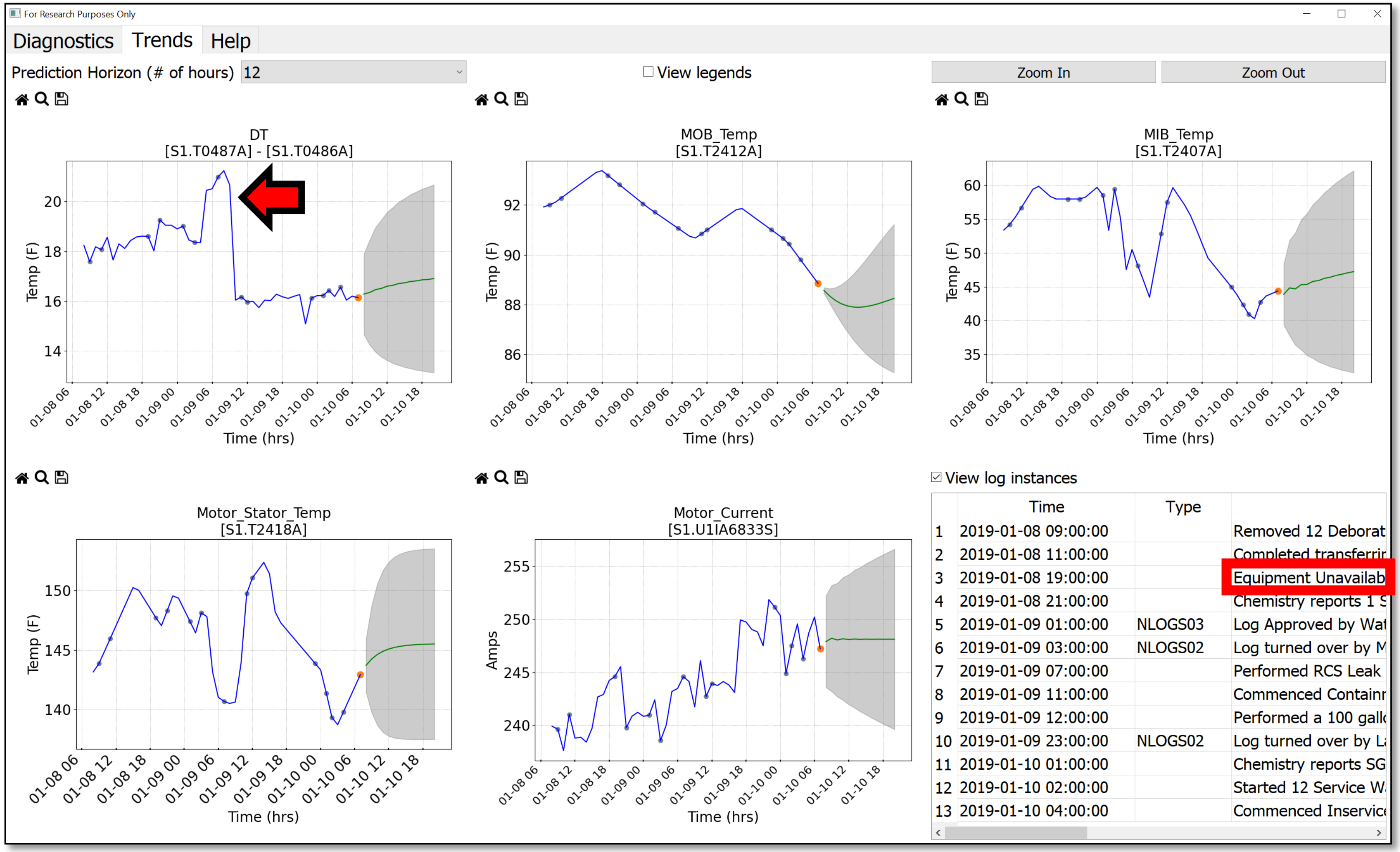
□ □ Motor_Current

□ Threshold

	Variable	Value	Unit
1	DT	16.136	F
2	Motor Current	247.21	Amps
3	MOB Temp	88.847	F
4	MIB Temp	44.369	F
5	Stator Temp	142.92	F
6	Gross Load	1219.9	MW

	ML Output	Diagnosis	Conf
1	Diagnosis	Waterbox Fouling	100.%
2	Inlier	No	N/A

Adding context to the data can further improve understanding.



Prediction Horizon (# of hours) 12

☐ View legends

☑ View log instances

	Time	Type	
1	2019-01-08 09:00:00		Removed 12 Deborat
2	2019-01-08 11:00:00		Completed transferrir
3	2019-01-08 19:00:00		Equipment Unavailab
4	2019-01-08 21:00:00		Chemistry reports 1 S
5	2019-01-09 01:00:00	NLOGS03	Log Approved by Wat
6	2019-01-09 03:00:00	NLOGS02	Log turned over by M
7	2019-01-09 07:00:00		Performed RCS Leak
8	2019-01-09 11:00:00		Commenced Containr
9	2019-01-09 12:00:00		Performed a 100 gallo
10	2019-01-09 23:00:00	NLOGS02	Log turned over by La
11	2019-01-10 01:00:00		Chemistry reports SG
12	2019-01-10 02:00:00		Started 12 Service W
13	2019-01-10 04:00:00		Commenced Inservice