Interpretable Machine Learning with R: LIME and SHAP

Hashai Papneja, Ph.D.
Texas A&M University – Central Texas

52nd SWDSI Conference (March 2023)
Houston, Texas

Agenda

1. The Concepts (~45 min)
   a) Interpretability
   b) Taxonomy of Interpretability Methods
   c) Local Interpretable Model-agnostic Explanations (LIME)
   d) SHapley Additive exPlanations (SHAP)

2. Hands-On Exercise in R (~45 min)

The Concepts

Interpretability

"Interpretability is the degree to which a human can understand the cause of a decision."

Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." arXiv Preprint arXiv:1706.07269. (2017).

"Interpretability is the degree to which a human can consistently predict the model's result."

Kim, Been, Rajiv Khanna, and Oluwasanmi O. Koyejo. "Examples are not enough, learn to criticize! Criticism for interpretability." Advances in Neural Information Processing Systems (2016).

Interpretability (contd.)

The easier it is for a human to understand why a decision or a prediction was made, the higher the interpretability of that Machine Learning (ML) model.

Interpretable Machine Learning (IML) can refer to the "extraction of relevant knowledge from a machine-learning model concerning relationships either contained in data or learned by the model".

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. "Definitions, methods, and applications in interpretable machine learning." Proceedings of the National Academy of Sciences, 116(44), 22071-22080. (2019).

Why Interpretability?

For certain problems or tasks it is not enough to get the prediction (the what). The model must also explain how it came to the prediction (the why), because a correct prediction only partially solves your original problem.

Doshi-Velez, Finale, and Been Kim. “Towards a rigorous science of interpretable machine learning,” no. Ml: 1–13. http://arxiv.org/abs/1702.08608 (2017).

Why Interpretability?

- Human curiosity and learning
- Finding meaning in the world
- Detecting bias
- Debugging and auditing ML models
- Increasing social acceptance of ML models

Doshi-Velez, Finale, and Been Kim. “Towards a rigorous science of interpretable machine learning,” no. Ml: 1–13. http://arxiv.org/abs/1702.08608 (2017).

Taxonomy of Interpretability Methods

- Intrinsic vs. Post Hoc Interpretability
- Intrinsic Interpretability
  - Refers to ML models that are considered interpretable due to their simple structure, such as short decision trees or sparse linear models.
- Post Hoc Interpretability
  - Refers to the application of interpretation methods after model training and outcome generation.

Taxonomy of Interpretability Methods

- Model-specific vs. Model-agnostic Methods
- Model-specific Methods
  - Are limited to specific model classes.
  - Usually look “within” the model.
- Model-agnostic Methods
  - Can be used on any machine learning model, and are applied after the model has been trained (post hoc).
  - Usually work by analyzing feature input and output pairs.
  - By definition, these methods do not access model internals such as weights or structural information.

Taxonomy of Interpretability Methods

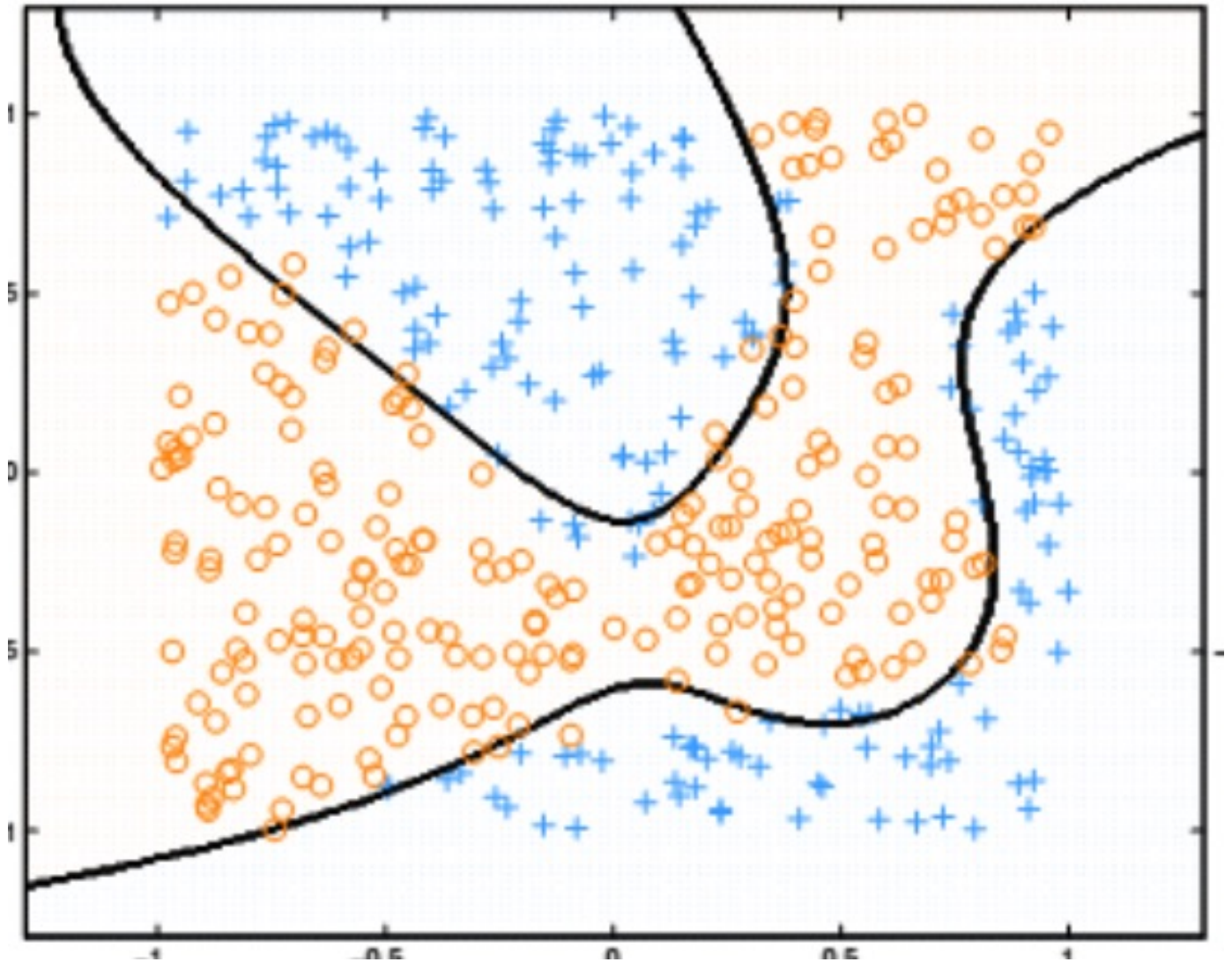
- Local vs. Global Explanation Methods
- Local
  - The interpretation method explains an individual prediction.
- Global
  - The interpretation method explains the behavior of the entire model.
  - Difficult to achieve in practice.

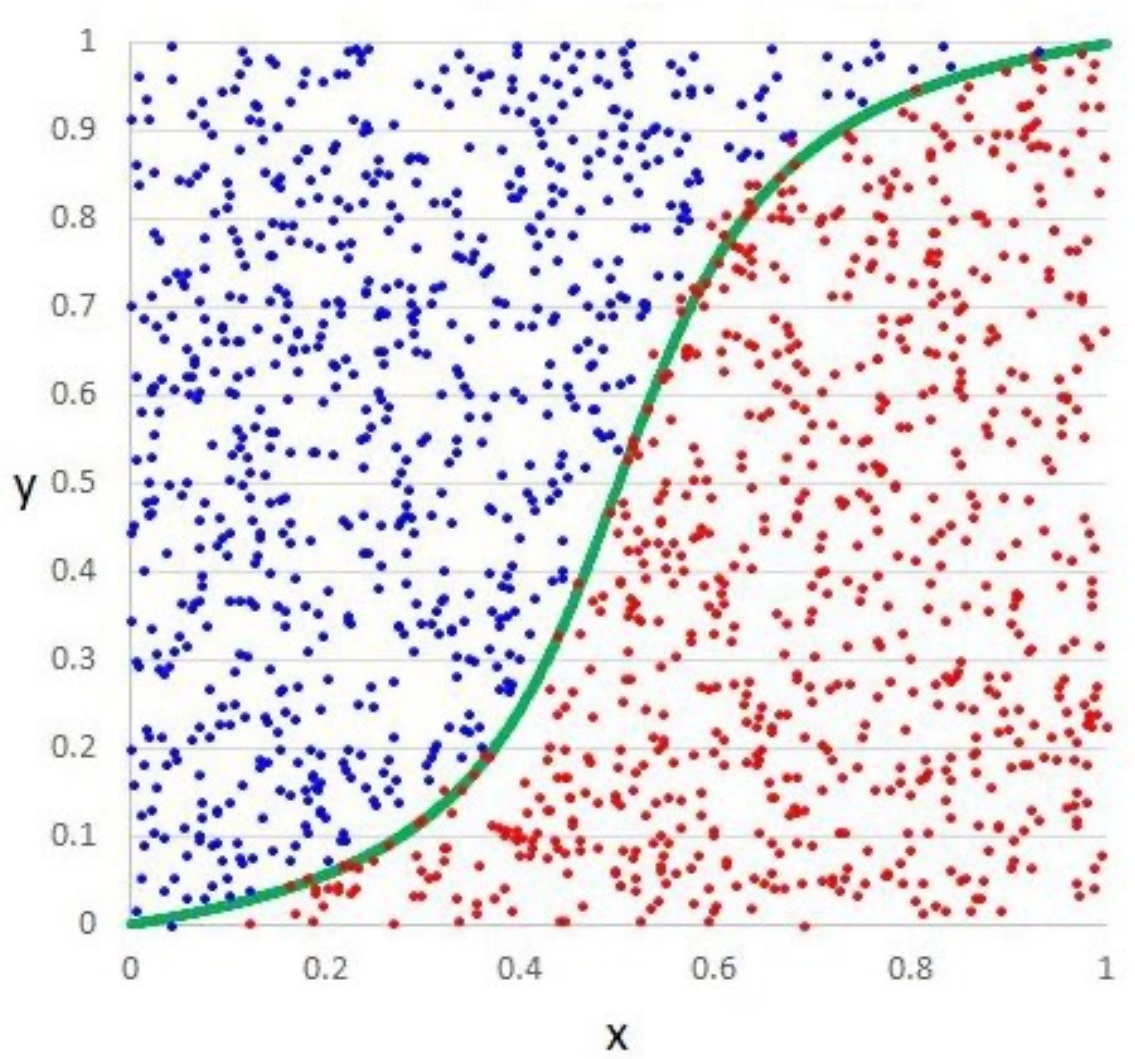
LIME and SHAP

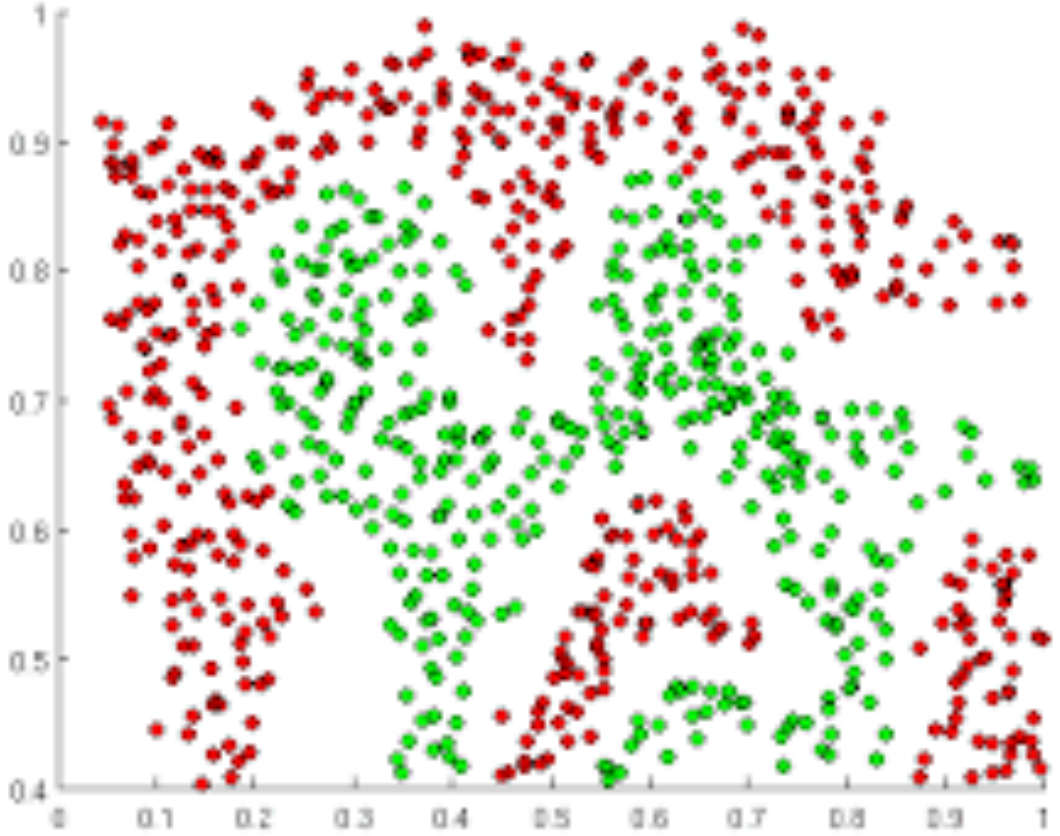
- LIME: Local Interpretable Model-agnostic Explanations
  - Post hoc
  - Local
  - Model-agnostic
- SHAP: SHapley Additive exPlanations
  - Post hoc
  - Local
  - Model-agnostic

LIME

- Goal: Understand why the ML model made a certain prediction.
- Key assumption: Every complex (i.e., black box) model is linear on a local scale.


1
5
0
5
1


1
0.9
0.8
0.7
0.6
y 0.5
0.4
0.3
0.2
0.1
0
0
0.2
0.4
0.6
0.8
1
x


1
0.9
0.8
0.7
0.6
0.5
0.4
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

LIME

- It is possible to fit a simple (i.e., an intrinsically explainable) model around a single observation that will mimic how the global model behaves at that locality.
- The simple model can then be used to explain the predictions of the more complex model locally.
  - LIME focuses on training local surrogate models to explain individual predictions.

LIME

- Select the instance of interest for which you want to have an explanation of the black box prediction.
- Perturb the dataset and get the black box predictions for these new points.
- Weight the new samples according to their proximity to the instance of interest.
- Train a weighted, interpretable model on the dataset with the variations.
- Explain the prediction by interpreting the local surrogate model.

SHAP

- A prediction can be explained by assuming that each feature value of the instance is a “player” in a game where the prediction is the payout.
- Shapley values – a method from coalitional game theory – tells us how to fairly distribute the “payout” among the features.
- The Shapley value, coined by Lloyd Shapley (1953), is a method for assigning payouts to players depending on their contribution to the total payout.
- The Shapley value is the average of all the marginal contributions to all possible coalitions.

Shapley, Lloyd S. “A value for n-person games.” Contributions to the Theory of Games 2.28 (1953): 307-317.

Lundberg, Scott M., and Su-In Lee. “A unified approach to interpreting model predictions.” Advances in Neural Information Processing Systems (2017).

SHAP

- 1, 2, 3, 4 → V
- $V_{1234} – V_{234} = MC_1$ to C_{234}
- $V_{123} – V_{23} = MC_1$ to C_{23}
- $V_{134} – V_{34} = MC_1$ to C_{34}
- …
- Average of all the marginal contributions to all possible coalitions = Shapley value of Member 1.

SHAP

- The interpretation of the Shapley value is: Given the current set of feature values, the Shapley value of a feature is the contribution of that feature to the difference between the actual prediction and the mean prediction.
- It is NOT the difference in the predicted value after removing the feature from model training.

Hands-on Exercise in R

Questions?

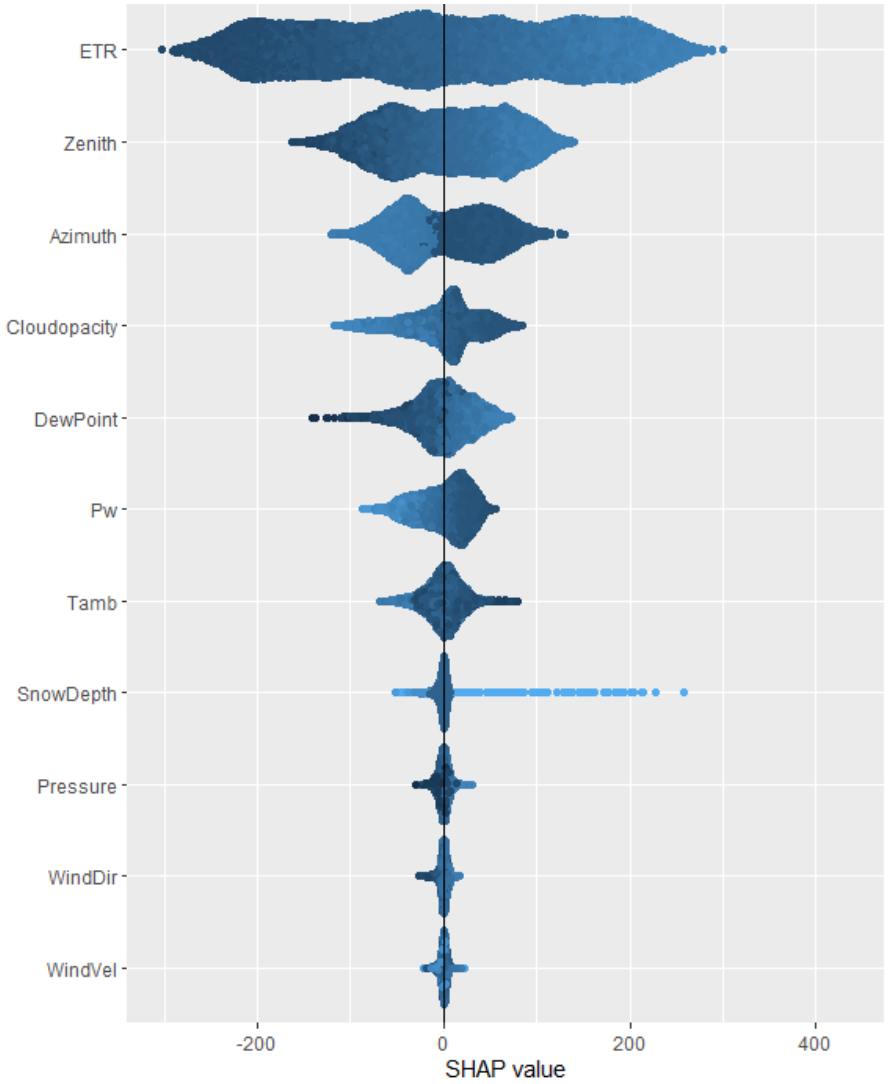
Thank You!

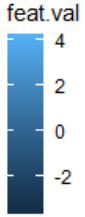
LIME – Disadvantages

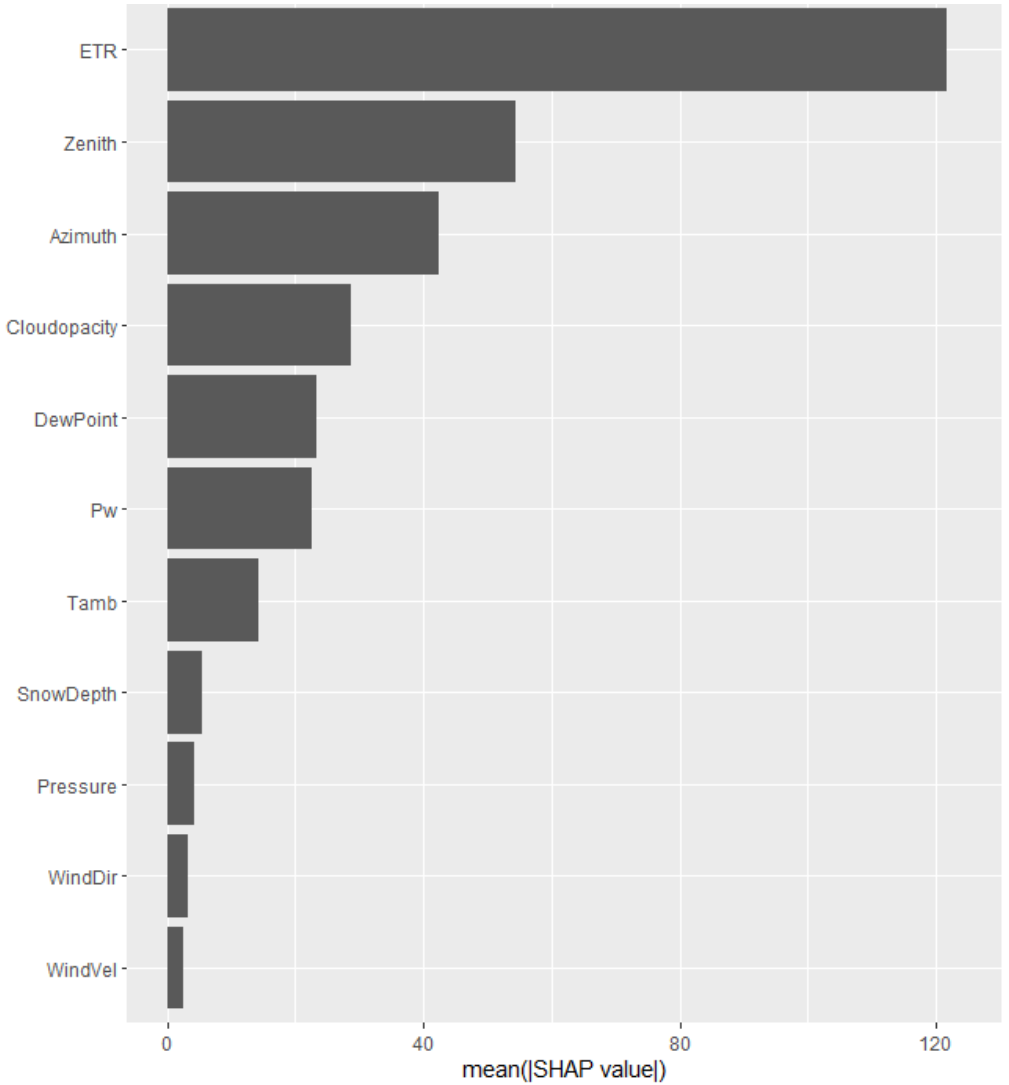
- What is a good kernel width?
- Sampling gives rise to unstable explanations.
- Can be manipulated to hide biases.

SHAP – Disadvantages

- Requires high computation time.
- Can be misinterpreted easily.
- Requires access to the data (in addition to access to the prediction function).


ETR
Zenith
Azimuth
Cloudopacity
DewPoint
Pw
Tamb
SnowDepth
Pressure
WindDir
WindVel
-200
0
200
400
SHAP value


feat.val
4
2
0
-2


ETR
Zenith
Azimuth
Cloudopacity
DewPoint
Pw
Tamb
SnowDepth
Pressure
WindDir
WindVel
0
40
80
120
mean(|SHAP value|)

Explanations

"An explanation is an assignment of causal responsibility" – Josephson and Josephson.

J.R. Josephson, S.G. Josephson
Abductive Inference: Computation, Philosophy, Technology
Cambridge University Press (1996)

"To explain an event is to provide some information about its causal history. In an act of explaining, someone who is in possession of some information about the causal history of some event – explanatory information, I shall call it – tries to convey it to someone else." – Lewis.

D. Lewis
Causal explanation
Philos. Pap., 2 (1986), pp. 214-240