Interpretable Machine Learning with R: PDP, ICE Plot, and Feature Importance Values

Hashai Papneja, PhD

53rd SWDSI Conference (April 2024)
Galveston, Texas

Agenda

1. The Concepts (~30 min)
   a) Basics of Interpretability – What, Why, and How
   b) Partial Dependence Plot (PDP)
   c) Individual Conditional Expectation (ICE) Plot
   d) Feature Importance Values

2. Hands-On Exercise in R (~60 min)

The Concepts

What is Interpretability?

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

Tim Miller. “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.”

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

What is Interpretability?

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).

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.

Why Interpretability?

- Human curiosity and learning
- Finding meaning in the world
- Debugging and auditing ML models
- Detecting bias
- Building trust and acceptance in 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).

How is Interpretability Implemented?

- 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.

How is Interpretability Implemented?

- Model-Specific vs. Model-Agnostic Methods
- Model-Specific Methods
  - Are limited to specific model classes. E.g., the interpretation of regression coefficients in a linear model.
  - Usually look “within” the model.

How is Interpretability Implemented?

- 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, or by creating a surrogate interpretable model.
  - By definition, these methods do not access model internals such as weights or structural information.

How is Interpretability Implemented?

- 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.

Interpretability Methods

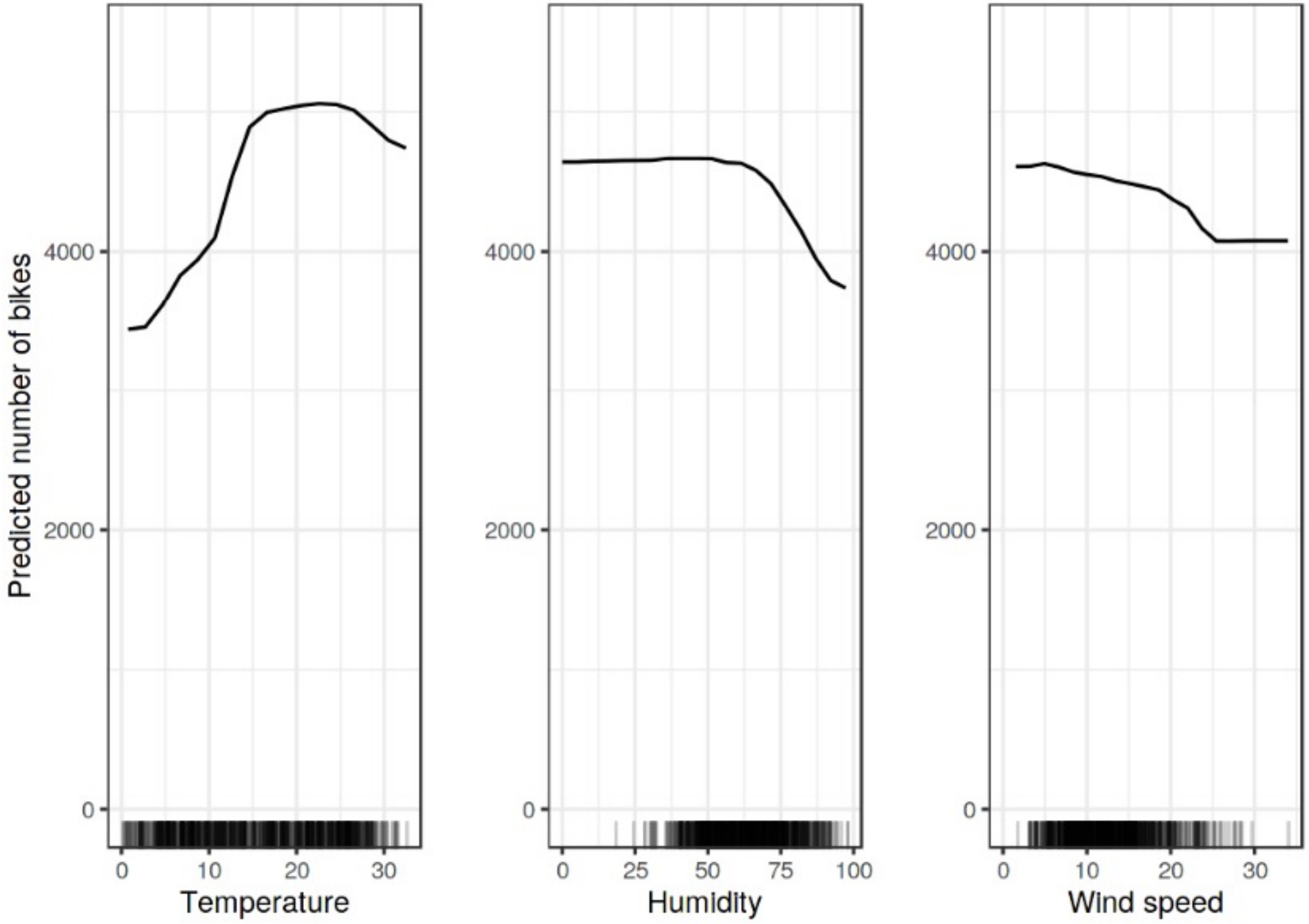
- Partial Dependence Plot (PDP)
- Individual Conditional Expectation (ICE) Plot
- Feature Importance Values

Partial Dependence Plot (PDP)

- The partial dependence plot (PDP or PD plot) shows the marginal effect that one or two features have on the predicted outcome of an ML model (Friedman 2001).

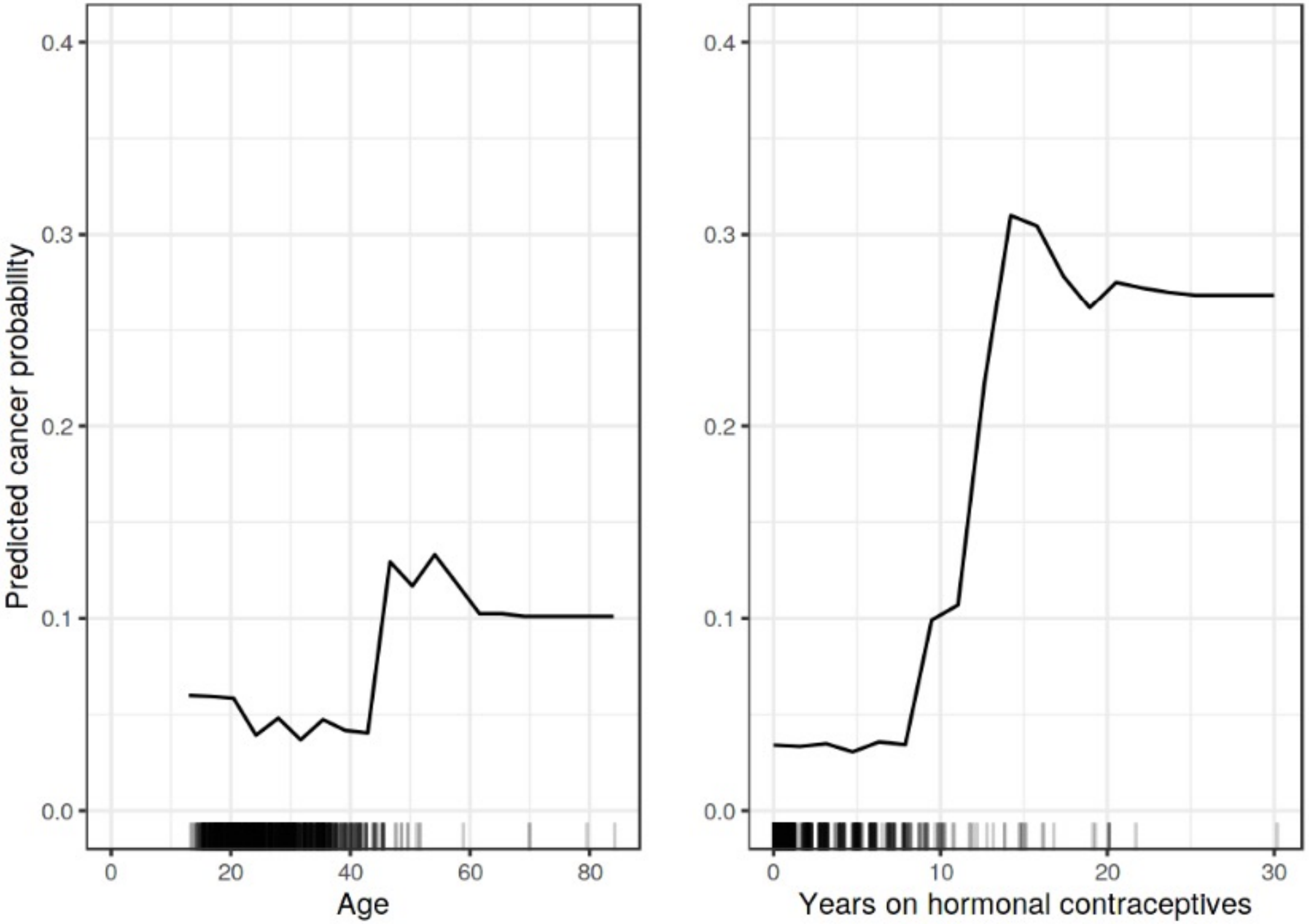
Friedman, Jerome H. “Greedy function approximation: A gradient boosting machine.” Annals of statistics (2001): 1189-1232.

Partial Dependence Plot (PDP)



Source: *Interpretable Machine Learning* by Chris Molnar

Partial Dependence Plot (PDP)



Source: *Interpretable Machine Learning* by Chris Molnar

Partial Dependence Plot (PDP)

- Replicate the dataset *m* times, where *m* is the number of unique values feature x_1 can take
  - Assign the same value to x_1 for all rows of replica *1*
  - Calculate the predicted outcome using the trained model *f*
  - Average the predicted outcomes
  - Plot the point
- Repeat for all unique values of x_1 (*m* times)
- What do we do for categorical features?

Partial Dependence Plot (PDP)

- For a classification model that outputs probabilities, the PDP displays the probability for a certain class given different values for the feature x_1.
- An assumption of the PDP is that the feature x_1 is not correlated with the other features.
    - If this assumption is violated, the averages calculated for the PDP will include data points that are very unlikely or even impossible in reality. (This is a disadvantage of PDPs.)

Partial Dependence Plot (PDP)

Advantages

- Model agnostic
- Intuitive
- Interpretation is clear
- Easy to implement

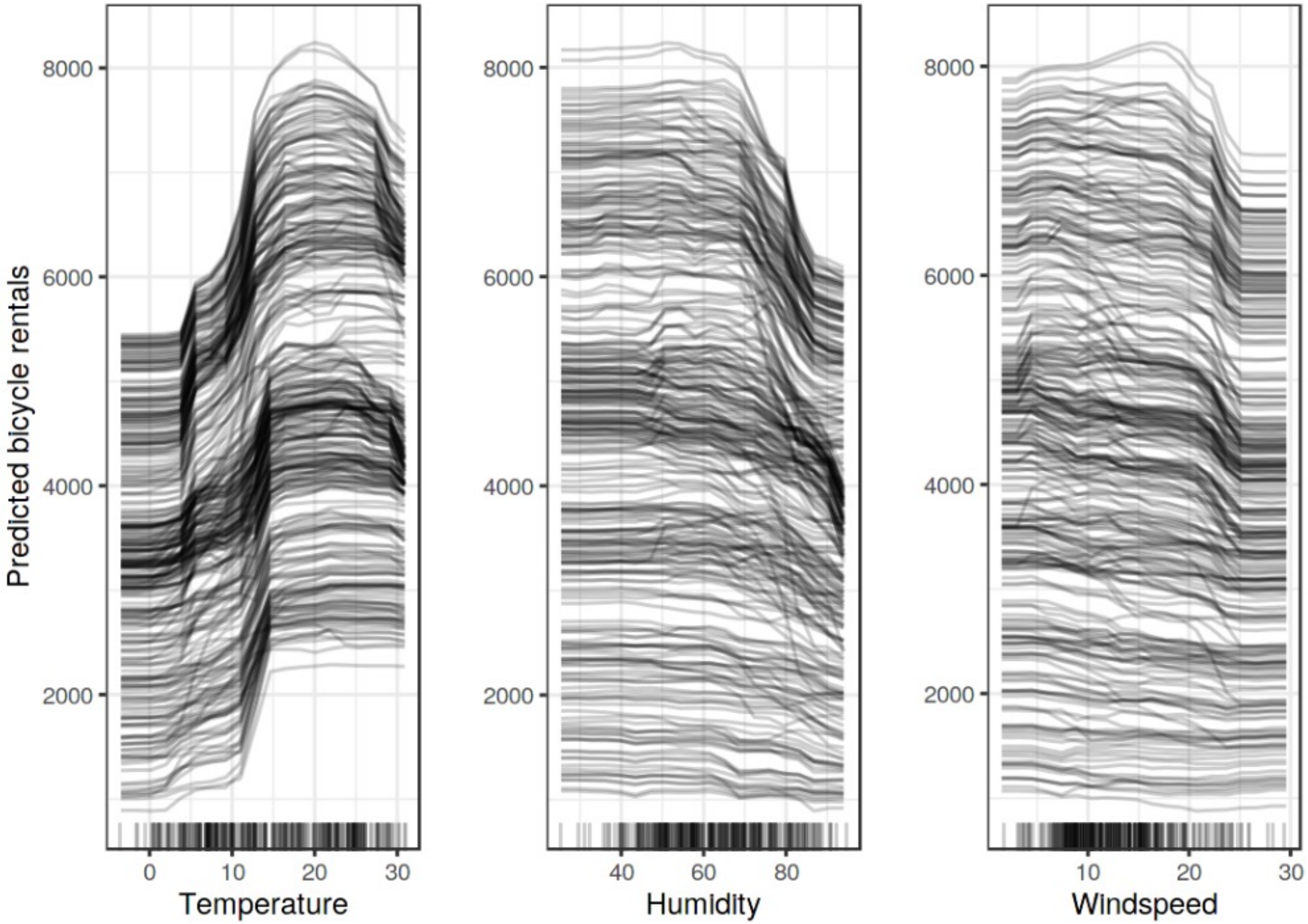
Disadvantages

- Assumption of independence
- Heterogenous effects might be hidden

Individual Conditional Expectation (ICE) Plot

- Display one line per instance that shows how the instance’s prediction changes when a feature changes.
- The PDP does not focus on specific instances, and hence gives the value for an average effect of the feature. It is a global method, but on an overall average.
- An ICE plot visualizes the dependence of the prediction on a feature for each instance separately, resulting in one line per instance (compared to one line overall in PDPs).
- A PDP is the average of the lines of an ICE plot.

Individual Conditional Expectation (ICE) Plot



Source: *Interpretable Machine Learning* by Chris Molnar

Individual Conditional Expectation (ICE) Plot

Advantages

- Model agnostic
- Intuitive
- Can uncover heterogeneous relationships

Disadvantages

- Can only display one feature meaningfully
- If the feature of interest is correlated with the other features, then some points in the lines might be invalid data points
- The plot can become overcrowded
- It might not be easy to see the average

Feature Importance Values

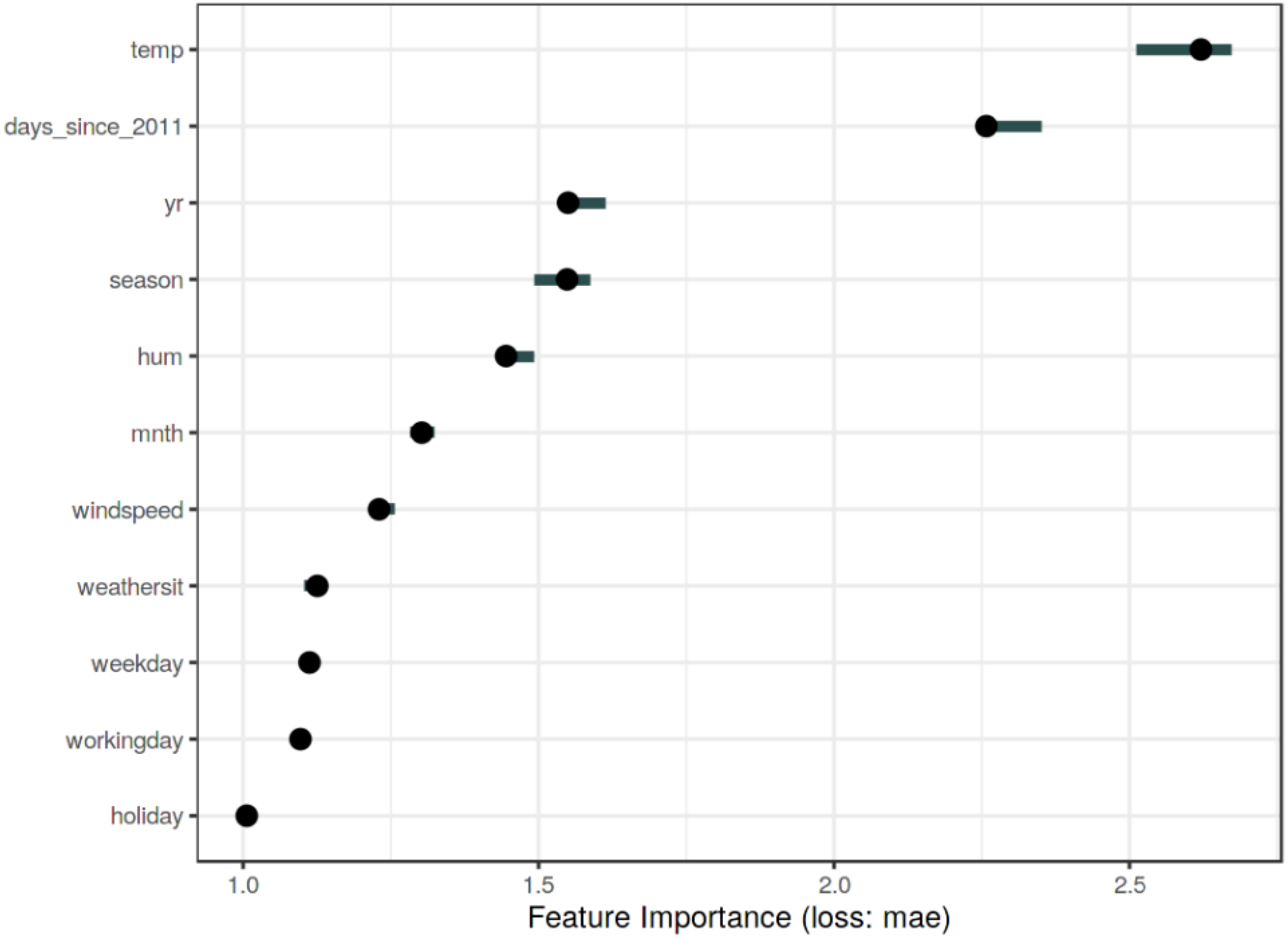
- The importance of a feature is the increase in the prediction error of the model after we permute the feature’s values (“permutation feature importance”).
- A feature is important if shuffling its values increases the model error, because the model relied on the feature for the prediction.
- A feature is unimportant if shuffling its values leaves the model error unchanged / slightly changed, because the model ignored the feature for the prediction.

Feature Importance Values

TEXAS A&M UNIVERSITY

- Start with the original dataset *X*, and calculate model performance
- Create a new dataset *X’* with values of the feature x_1 shuffled (permuted)
- Calculate model performance
- Calculate the change in model performance
  - This gives a quantified measure of how important the feature is.
- Repeat for all features

Feature Importance Values



Source: *Interpretable Machine Learning* by Chris Molnar

Feature Importance Values

Advantages

- Model agnostic
- Intuitive
- Provides a succinct, global insight into the model’s behavior

Disadvantages

- If features are correlated, the permutation feature importance can be biased by unrealistic data instances
- When the permutation is repeated, the results might vary greatly
- You need access to the true outcome to calculate error
- Feature importance is for that particular model, i.e., may differ for another model

https://tinyurl.com/swdsi24

Hands-on Exercise in R

Questions?

Thank You!