Interpretable Machine Learning (part 2)

April 2, 2026

Announcements

- **No office hours tomorrow**
- **Final Project Checkpoint 2 due Tuesday April 7 at 11:59pm**
  - Push all of the code that you have written so far to GitHub for your first code review. This code should be written in your `final project/` folder and organized using a file structure similar to your labs.
  - *Tip:* If you are completing a real data analysis project, it is recommended to have conducted a preliminary exploratory data analysis and modeling pass through the data. This would allow for more constructive feedback at this checkpoint.
  - **If you want more helpful feedback, leave comments describing what you are trying to do throughout the code and list your planned/future to-dos**

Plan for this week: Interpretability/Explainability

1. **Why Interpretability/Explainability?**
2. **What do we mean by Interpretable/Explainable ML/AI**
3. **Interpretability/Explainability Tools**

Last Time: Intro to Interpretability

Why do we care about interpretability?

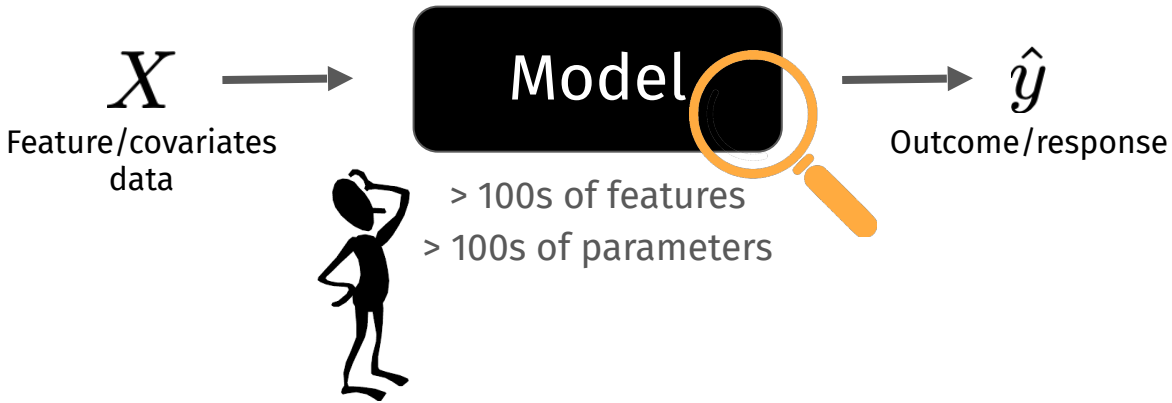
- Facilitates **scientific understanding** and decision-making
- Instills **trust/distrust** in a model
- Aids **human-in-the-loop** workflow
- Facilitates auditing for errors and **biases**

How can we interpret our machine learning models?

Avenue 1: "Let's develop interpretable* models from the start"

Can be visualized.
Implemented by hand.
Mirrors way people think.

Avenue 2: "Let's develop tools to interpret complex, black-box models"



* Many technical definitions... (e.g., Murdoch et al. (2019))

Overview of Interpretability Methods

Feature Importances
How important is a feature (or set of features) in making the model predictions?

Sample Influences
How does leaving out a sample change the model fit, predictions, and/or error?

Overview of Interpretability Methods

Feature Importances

How important is a feature (or set of features) in making the model predictions?

Sample Influences

How does leaving out a sample change the model fit, predictions, and/or error?

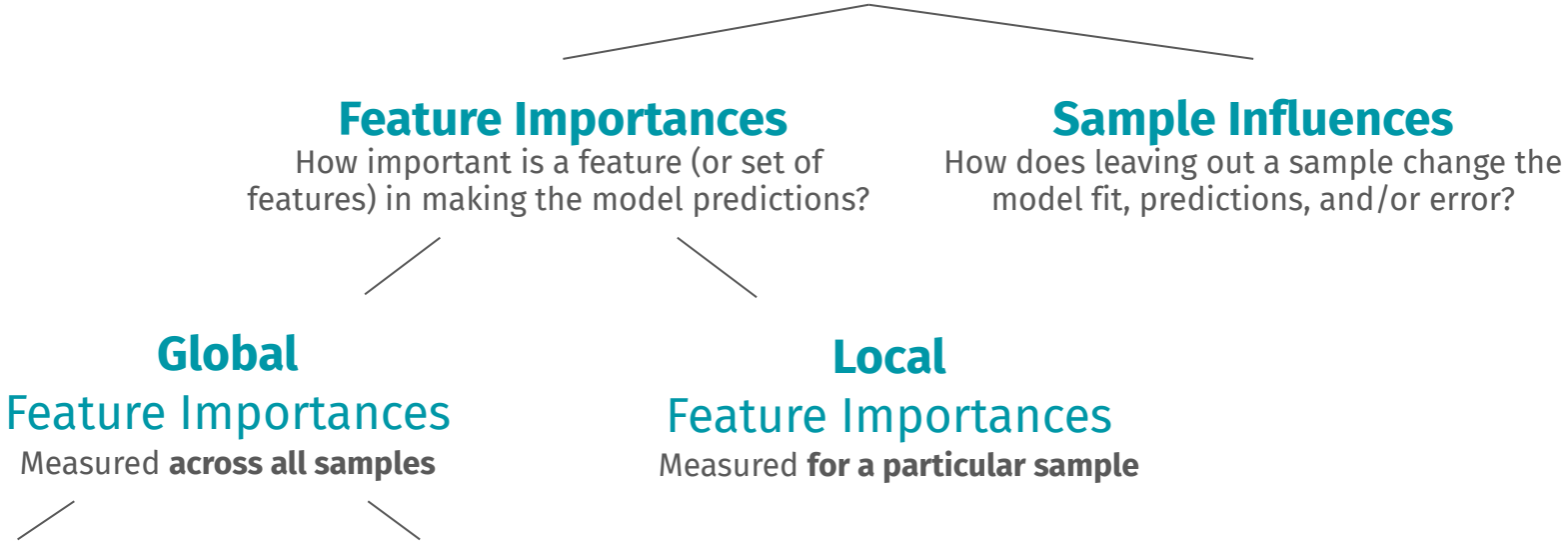
Global Feature Importances

Measured **across all samples**

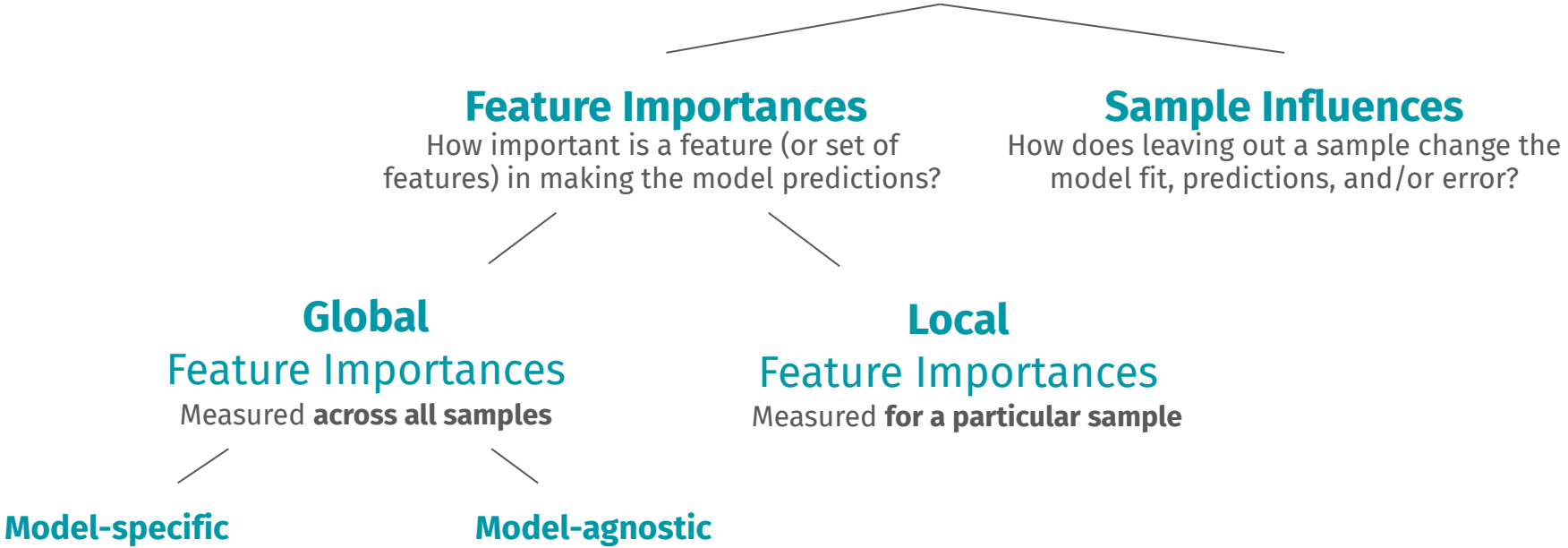
Local Feature Importances

Measured **for a particular sample**

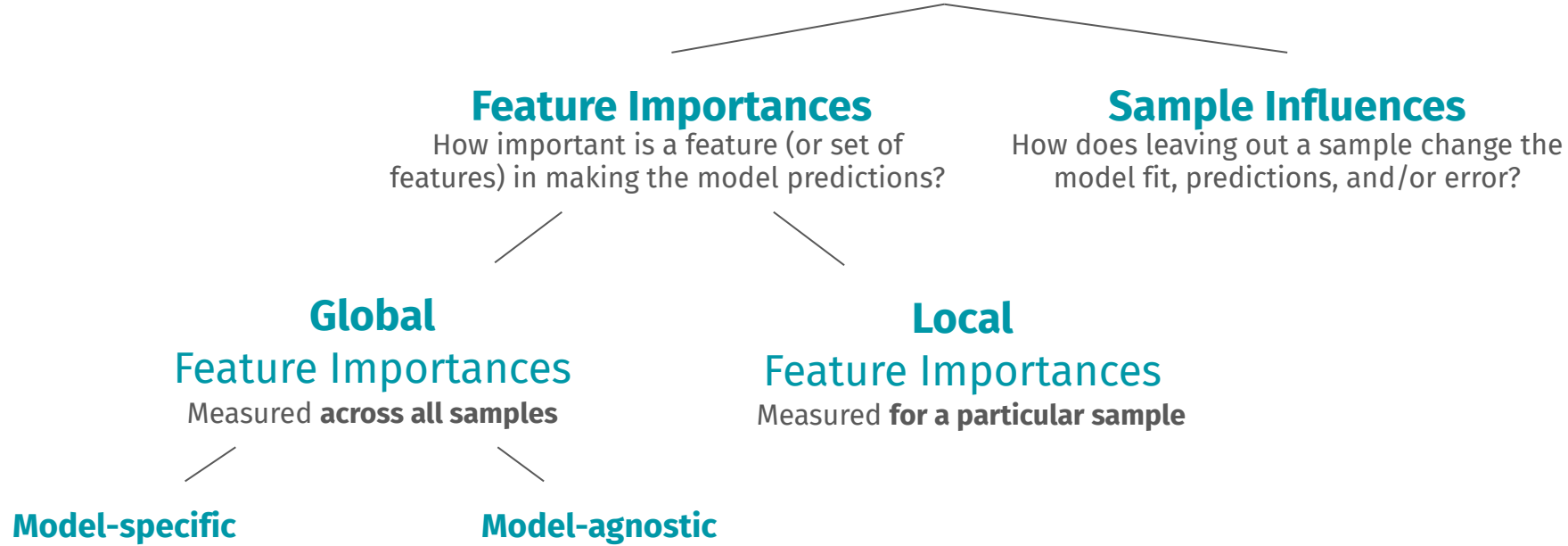
Overview of Interpretability Methods



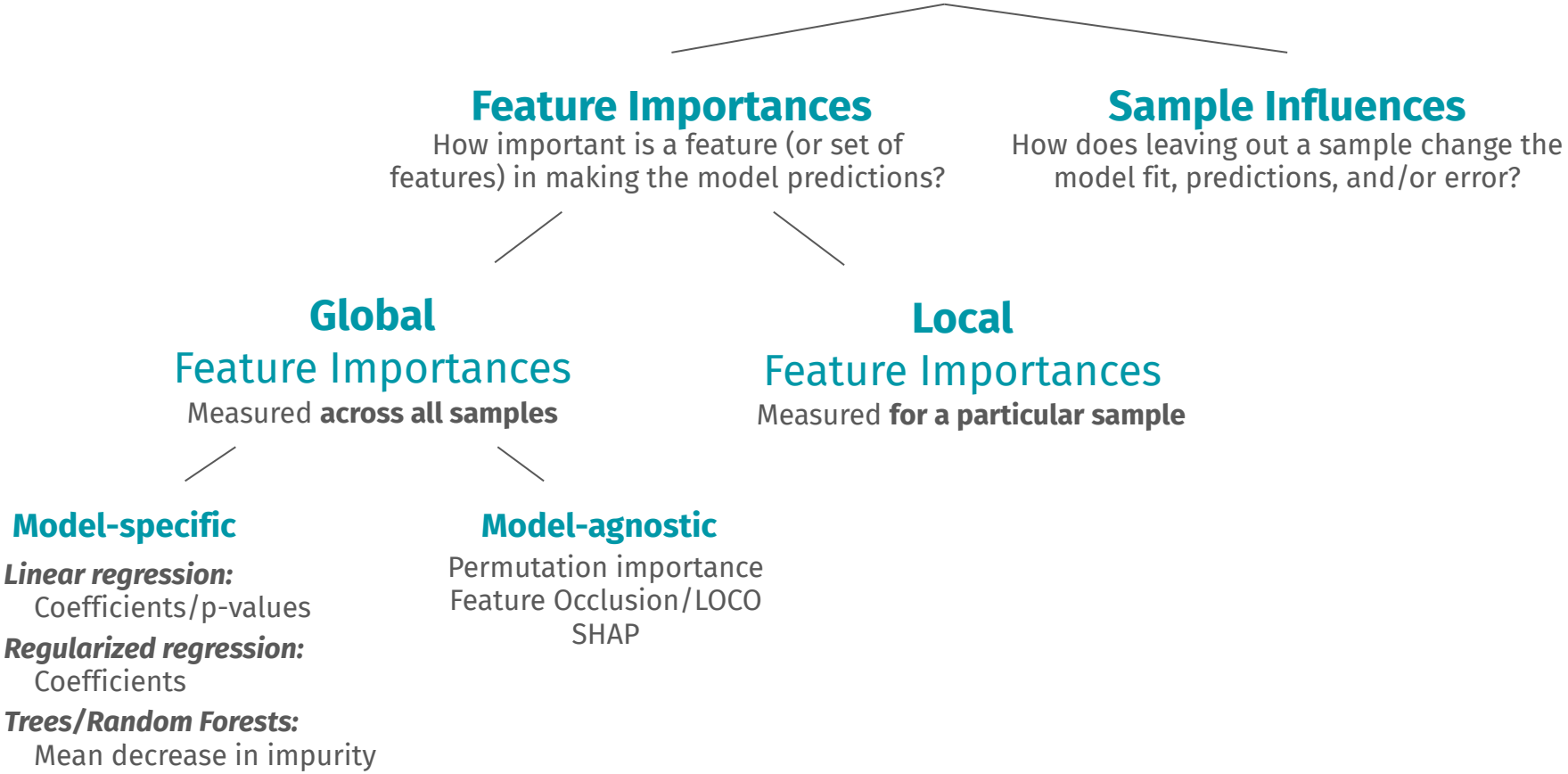
Overview of Interpretability Methods



Overview of Interpretability Methods



Overview of Interpretability Methods



Overview of Interpretability Methods

Feature Importances

How important is a feature (or set of features) in making the model predictions?

Sample Influences

How does leaving out a sample change the model fit, predictions, and/or error?

Global Feature Importances

Measured **across all samples**

Local Feature Importances

Measured **for a particular sample**

Model-specific

Linear regression:
Coefficients/p-values

Regularized regression:
Coefficients

Trees/Random Forests:
Mean decrease in impurity

Model-agnostic

Permutation importance
Feature Occlusion/LOCO
SHAP

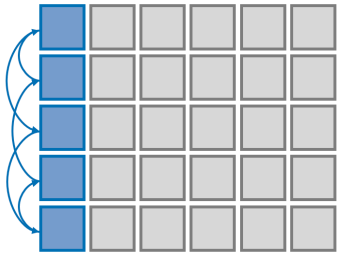
* Partial Dependence Plots

Feature **permutation** versus feature **occlusion** importance

To measure importance of feature *k*,

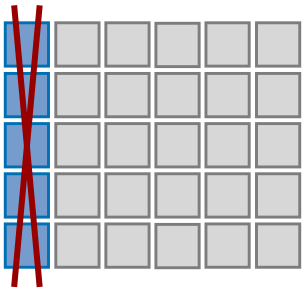
Feature Permutation

- Measures change in model's prediction error from **permuting** feature *k*
- No need to refit model

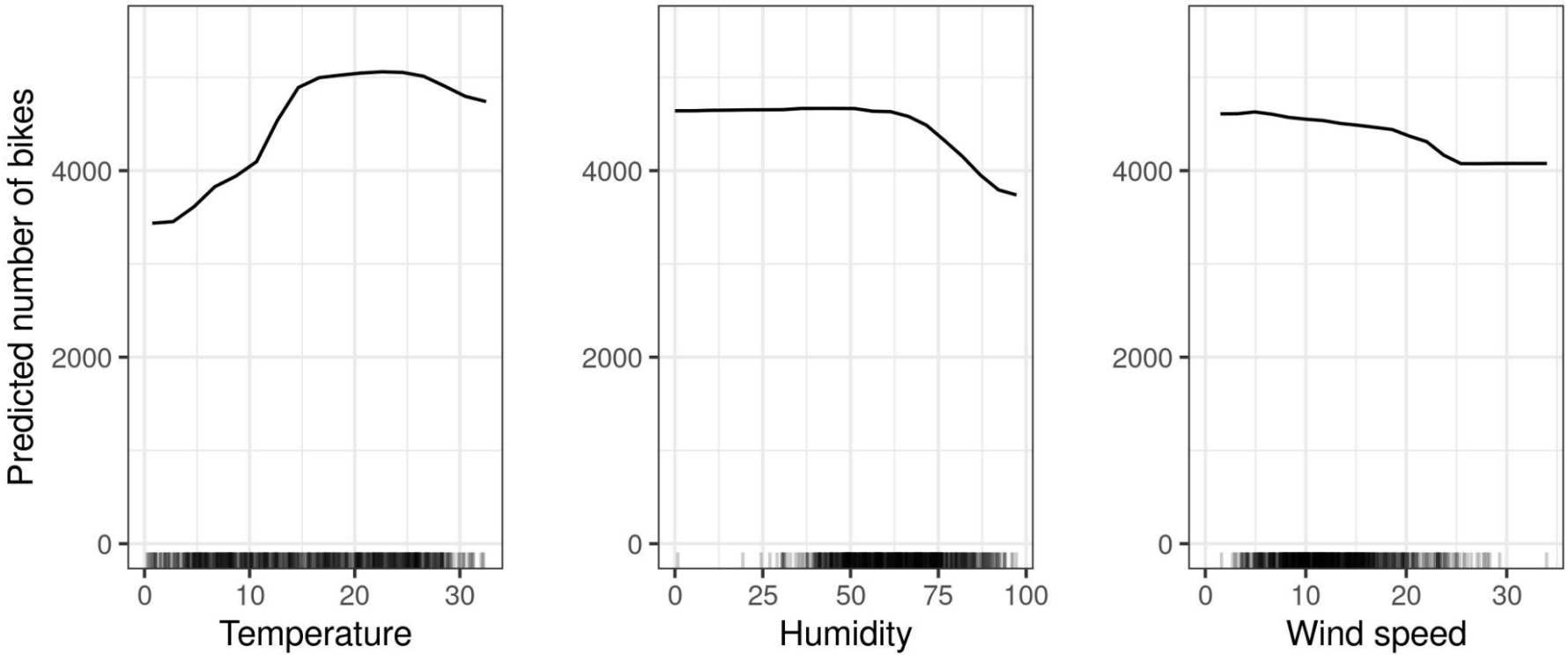


Feature Occlusion

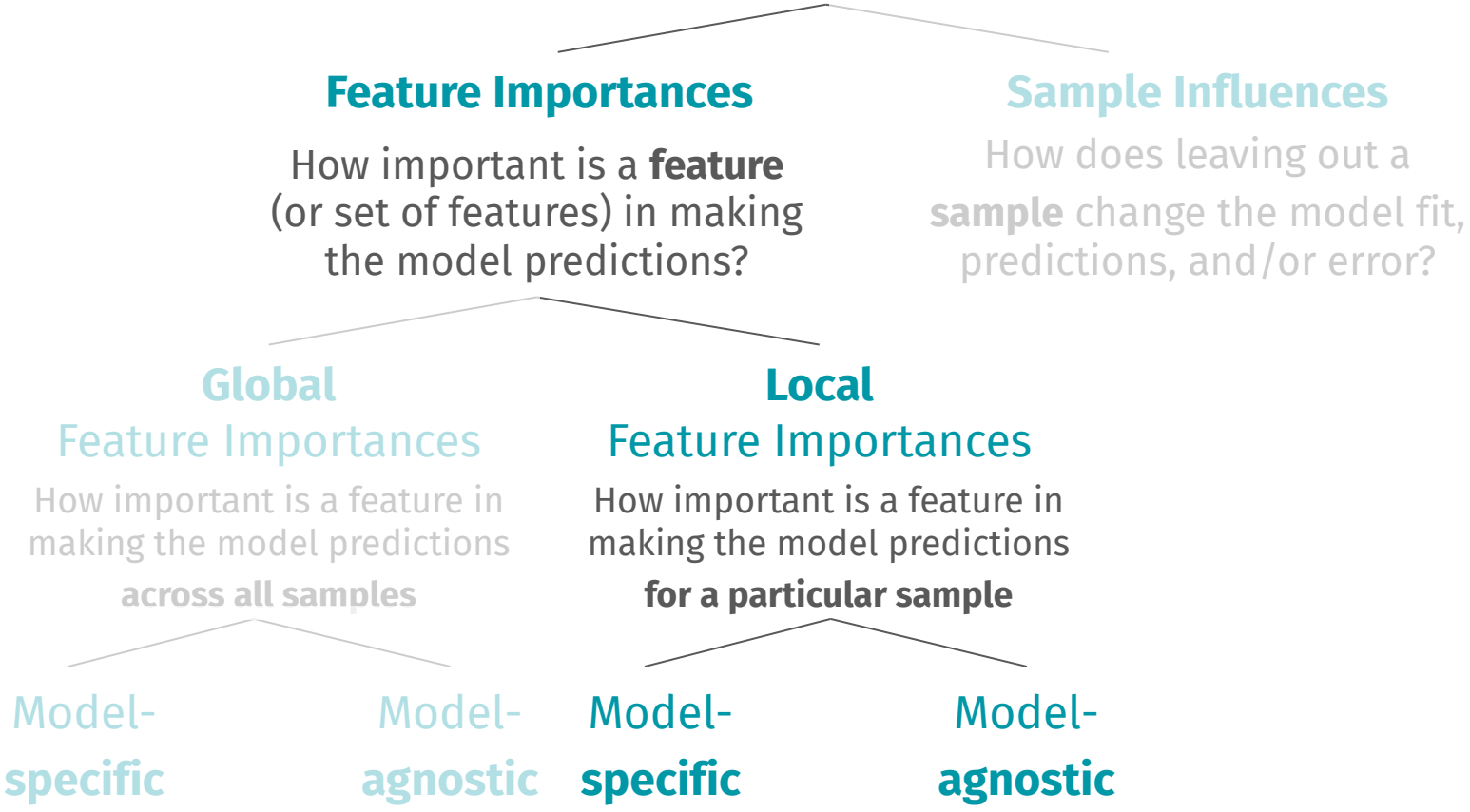
- Measures change in model's prediction error after refitting model **without** feature *k*
- Need to refit model for every *k*



Partial Dependence Plots (PDP)



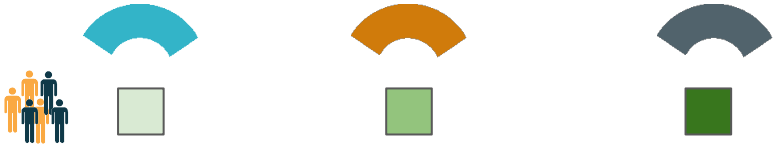
(One possible) Taxonomy of interpretations



Local Feature Importances

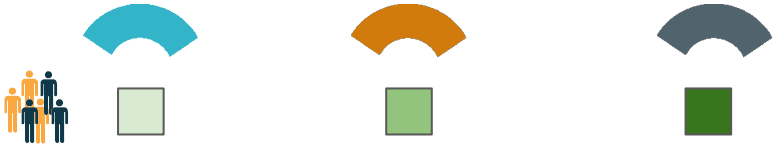
Global vs Local Feature Importances

"Global importance"

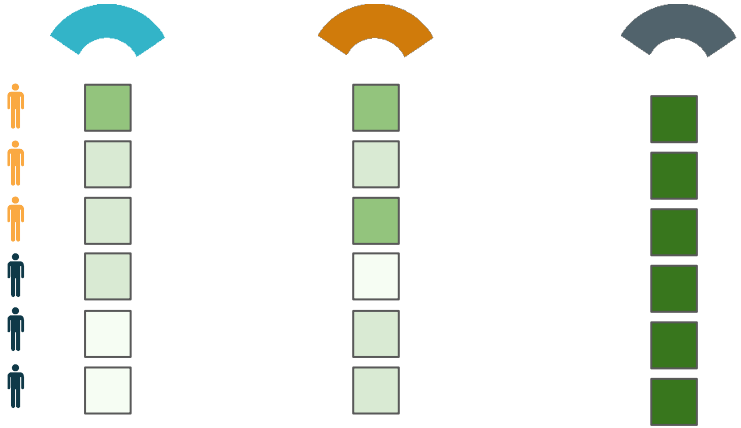


Global vs Local Feature Importances

"Global importance"

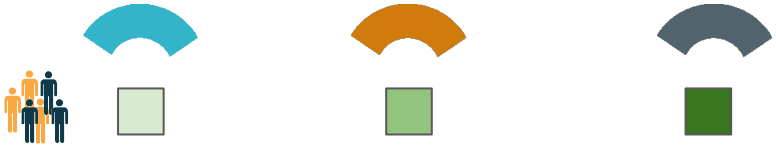


"Local importance" *(per-individual)*

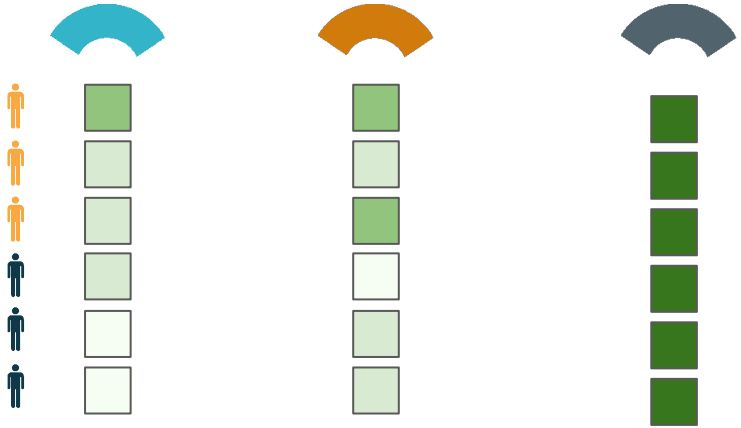


Global vs Local Feature Importances

"Global importance"



"Local importance" *(per-individual)*



Applications of local feature importances

- \+ Precision medicine
- \+ Personalized recommendations
- \+ Subgroup identification

LIME: Local Interpretable Model-agnostic Explanations

Main Idea: approximate the black-box model *locally* with an interpretable surrogate model and interpret the surrogate model

LIME: Local Interpretable Model-agnostic Explanations

Main Idea: approximate the black-box model *locally* with an interpretable surrogate model and interpret the surrogate model

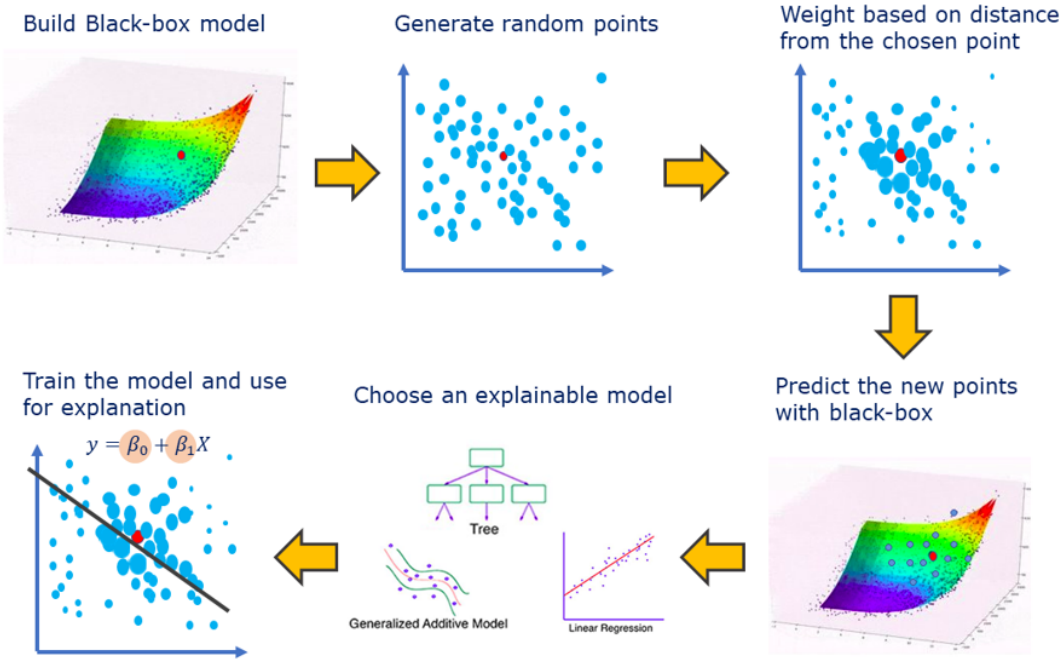
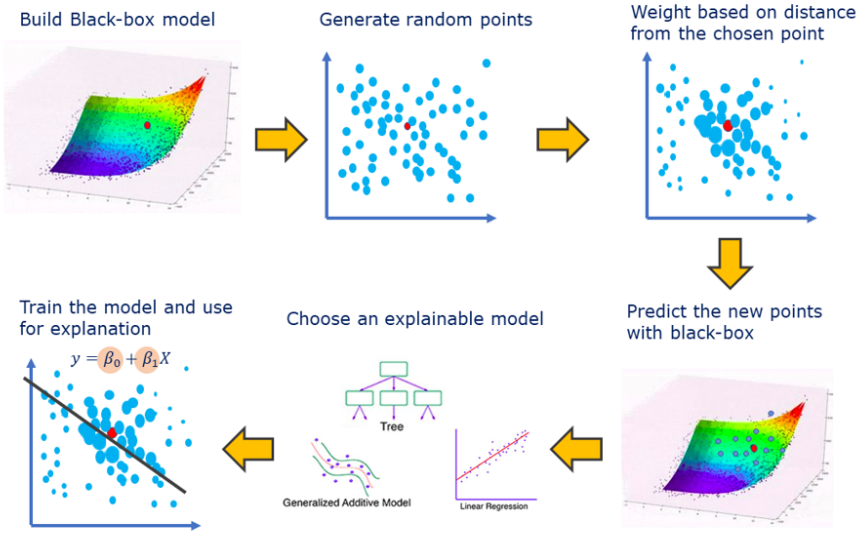


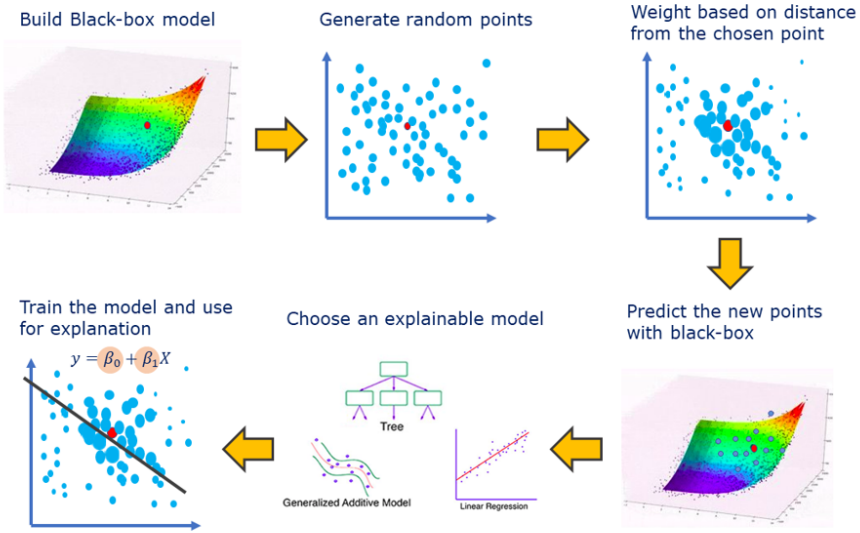
Image source: https://www.crif.digital/blog/lime-explain-machine-learning-predictions/

LIME: Local Interpretable Model-agnostic Explanations



LIME: Local Interpretable Model-agnostic Explanations

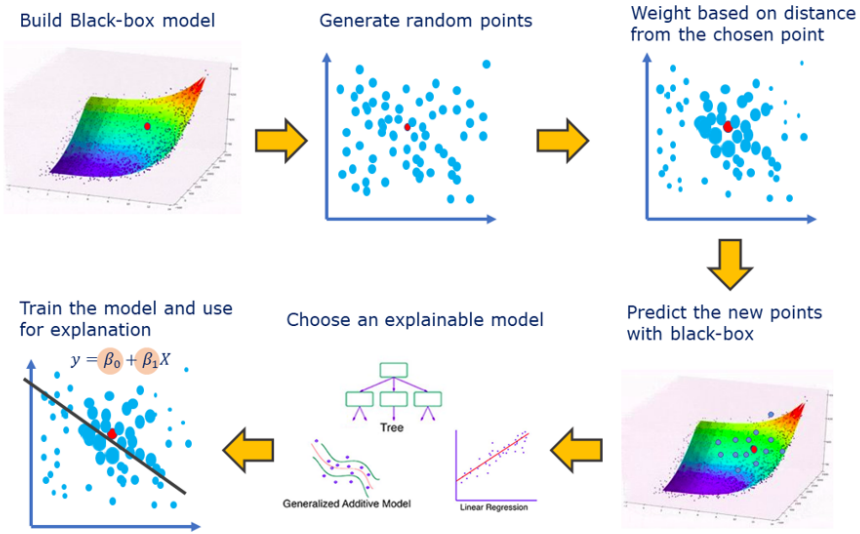
Practical Considerations/Choices:



LIME: Local Interpretable Model-agnostic Explanations

Practical Considerations/Choices:

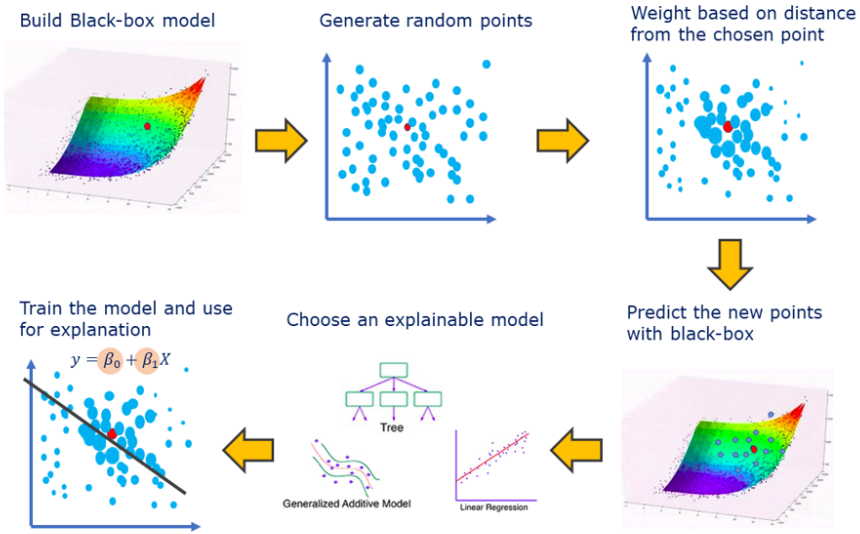
- How to generate random “close” samples?



LIME: Local Interpretable Model-agnostic Explanations

Practical Considerations/Choices:

- How to generate random “close” samples?
- How to weight samples? Typically using exponential kernel

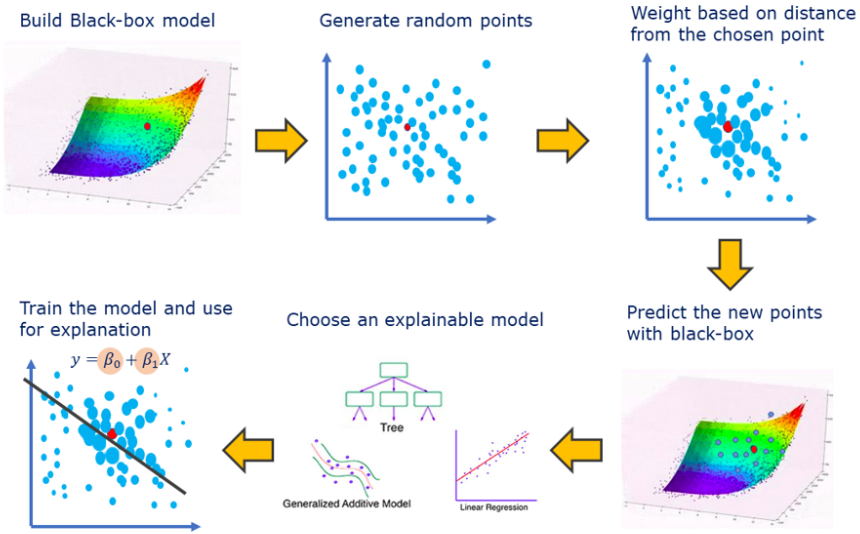


LIME: Local Interpretable Model-agnostic Explanations

Practical Considerations/Choices:

- How to generate random “close” samples?
- How to weight samples? Typically using exponential kernel

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right)$$



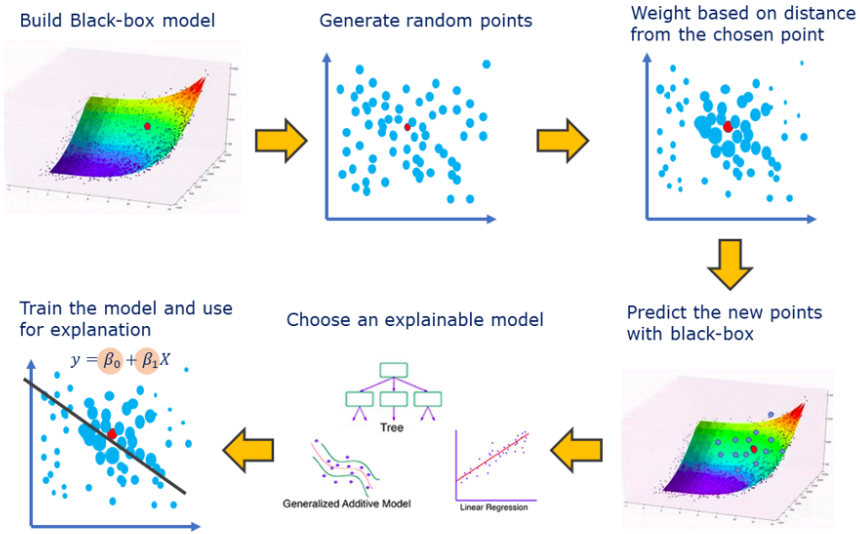
LIME: Local Interpretable Model-agnostic Explanations

Practical Considerations/Choices:

- How to generate random “close” samples?
- How to weight samples? Typically using exponential kernel

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right)$$

  - How to choose bandwidth σ of exponential kernel?



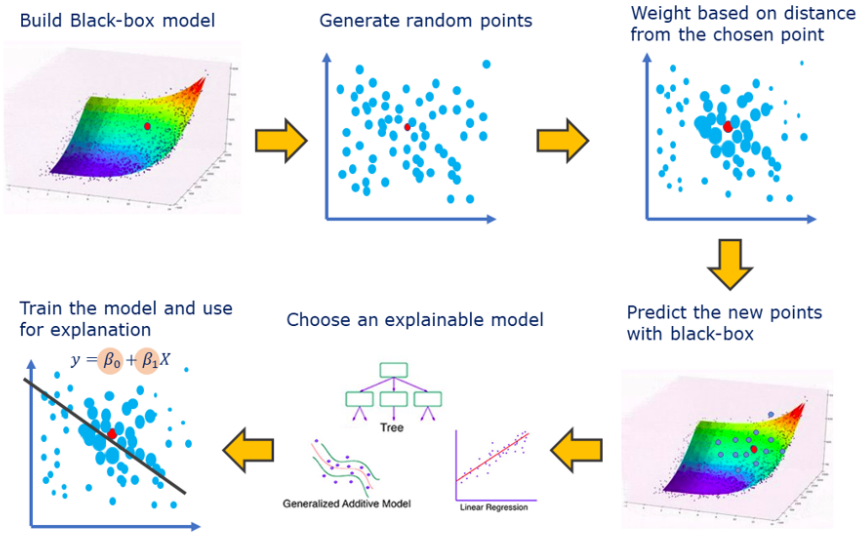
LIME: Local Interpretable Model-agnostic Explanations

Practical Considerations/Choices:

- How to generate random “close” samples?
- How to weight samples? Typically using exponential kernel

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right)$$

  - How to choose bandwidth σ of exponential kernel?
- Should we weight based upon euclidean distance to original training point if only a subset of variables are “important”?



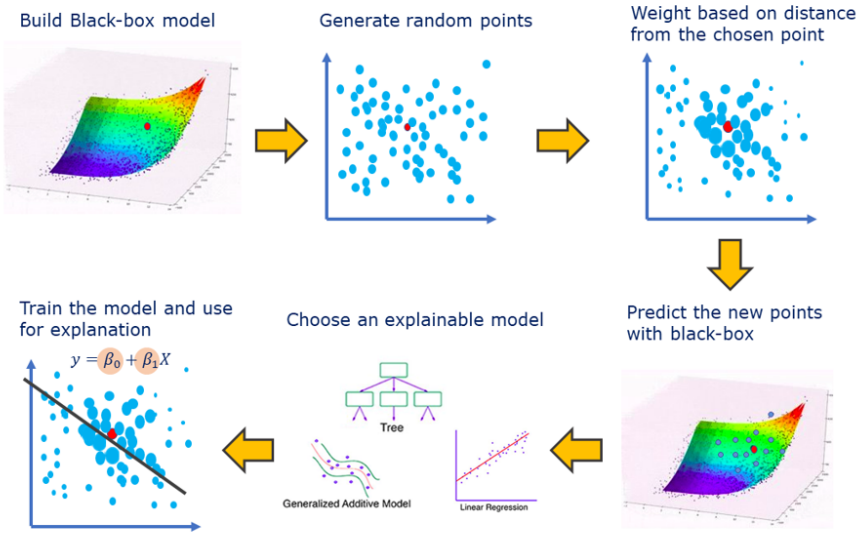
LIME: Local Interpretable Model-agnostic Explanations

Practical Considerations/Choices:

- How to generate random “close” samples?
- How to weight samples? Typically using exponential kernel

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right)$$

  - How to choose bandwidth σ of exponential kernel?
- Should we weight based upon euclidean distance to original training point if only a subset of variables are “important”?
- Which interpretable model to use?



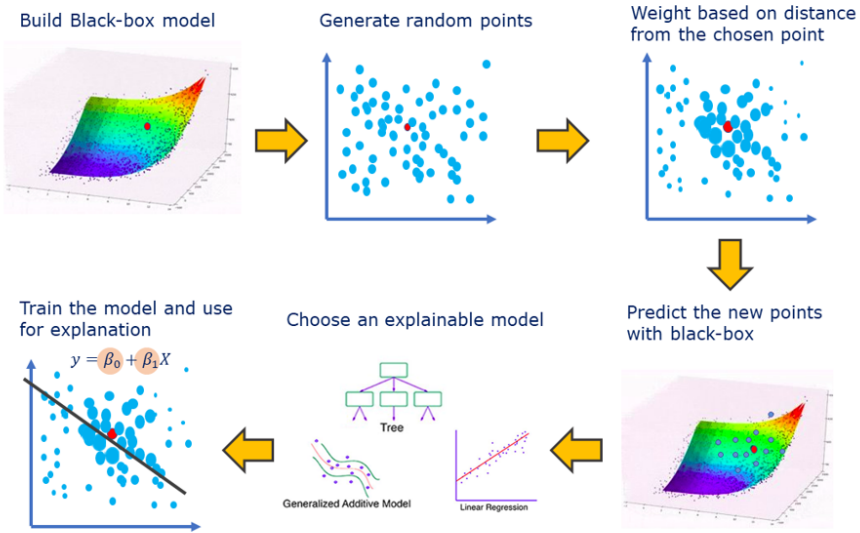
LIME: Local Interpretable Model-agnostic Explanations

Practical Considerations/Choices:

- How to generate random “close” samples?
- How to weight samples? Typically using exponential kernel

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right)$$

  - How to choose bandwidth σ of exponential kernel?
- Should we weight based upon euclidean distance to original training point if only a subset of variables are “important”?
- Which interpretable model to use?



Warning: known to be very unstable*, so use with caution

* Alvarez-Melis and Jaakkola (2018). On the robustness of interpretability methods.

Shapley Values in Game Theory

Shapley Values in Game Theory

- **Shapley Values:** a method from coalitional game theory, which measures the *average marginal contribution* of a feature value across all possible *coalitions*

Shapley Values in Game Theory

- **Shapley Values:** a method from coalitional game theory, which measures the *average marginal contribution* of a feature value across all possible *coalitions*
- **Idea:** Suppose we wanted to measure the importance of the blue person

Shapley Values in Game Theory

- **Shapley Values:** a method from coalitional game theory, which measures the *average marginal contribution* of a feature value across all possible *coalitions*
- **Idea:** Suppose we wanted to measure the importance of the blue person

"All possible coalitions"

Shapley Values in Game Theory

- **Shapley Values:** a method from coalitional game theory, which measures the *average marginal contribution* of a feature value across all possible *coalitions*
- **Idea:** Suppose we wanted to measure the importance of the blue person

"All possible coalitions"

Shapley Values in Game Theory

- **Shapley Values:** a method from coalitional game theory, which measures the *average marginal contribution* of a feature value across all possible *coalitions*
- **Idea:** Suppose we wanted to measure the importance of the blue person

"All possible coalitions"

Shapley Values in Game Theory

- **Shapley Values:** a method from coalitional game theory, which measures the *average marginal contribution* of a feature value across all possible *coalitions*
- **Idea:** Suppose we wanted to measure the importance of the blue person

"All possible coalitions"

Shapley Values in Game Theory

- **Shapley Values:** a method from coalitional game theory, which measures the *average marginal contribution* of a feature value across all possible *coalitions*
- **Idea:** Suppose we wanted to measure the importance of the blue person

"All possible coalitions"

Shapley Values in Game Theory

- **Shapley Values:** a method from coalitional game theory, which measures the *average marginal contribution* of a feature value across all possible *coalitions*
- **Idea:** Suppose we wanted to measure the importance of the blue person

"All possible coalitions"

Shapley Values in Game Theory

- **Shapley Values:** a method from coalitional game theory, which measures the *average marginal contribution* of a feature value across all possible *coalitions*
- **Idea:** Suppose we wanted to measure the importance of the blue person

"All possible coalitions"

Shapley Values in Game Theory

- **Shapley Values:** a method from coalitional game theory, which measures the *average marginal contribution* of a feature value across all possible *coalitions*
- **Idea:** Suppose we wanted to measure the importance of the blue person

"All possible coalitions"

Shapley Values in Game Theory

- **Shapley Values:** a method from coalitional game theory, which measures the *average marginal contribution* of a feature value across all possible *coalitions*
- **Idea:** Suppose we wanted to measure the importance of the blue person

"All possible coalitions"

vs

Shapley Values in Game Theory

- **Shapley Values:** a method from coalitional game theory, which measures the *average marginal contribution* of a feature value across all possible *coalitions*
- **Idea:** Suppose we wanted to measure the importance of the blue person

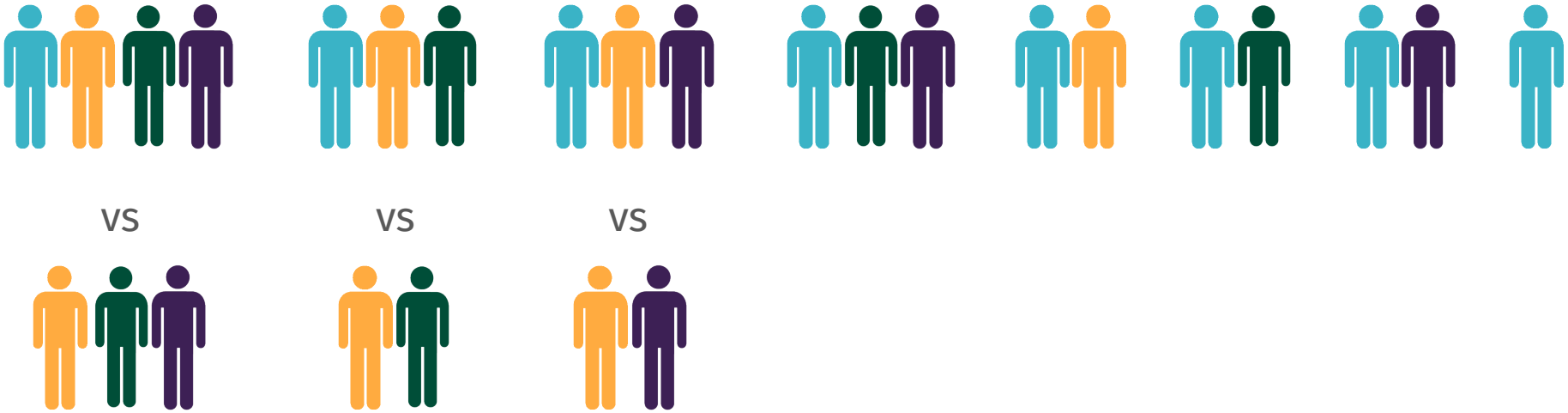
"All possible coalitions"

vs vs

Shapley Values in Game Theory

- **Shapley Values:** a method from coalitional game theory, which measures the *average marginal contribution* of a feature value across all possible *coalitions*
- **Idea:** Suppose we wanted to measure the importance of the blue person

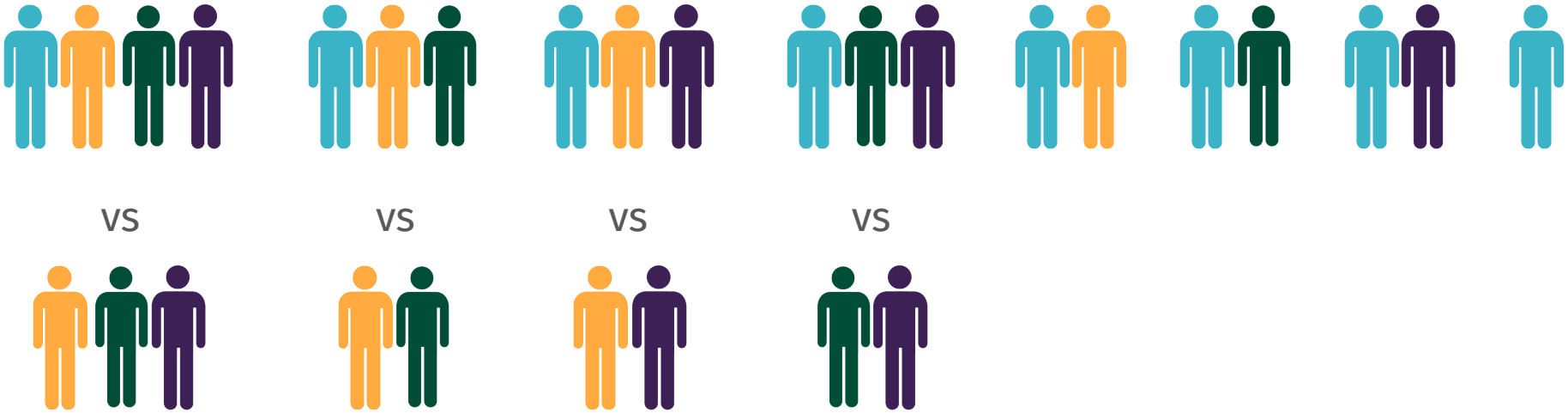
"All possible coalitions"



Shapley Values in Game Theory

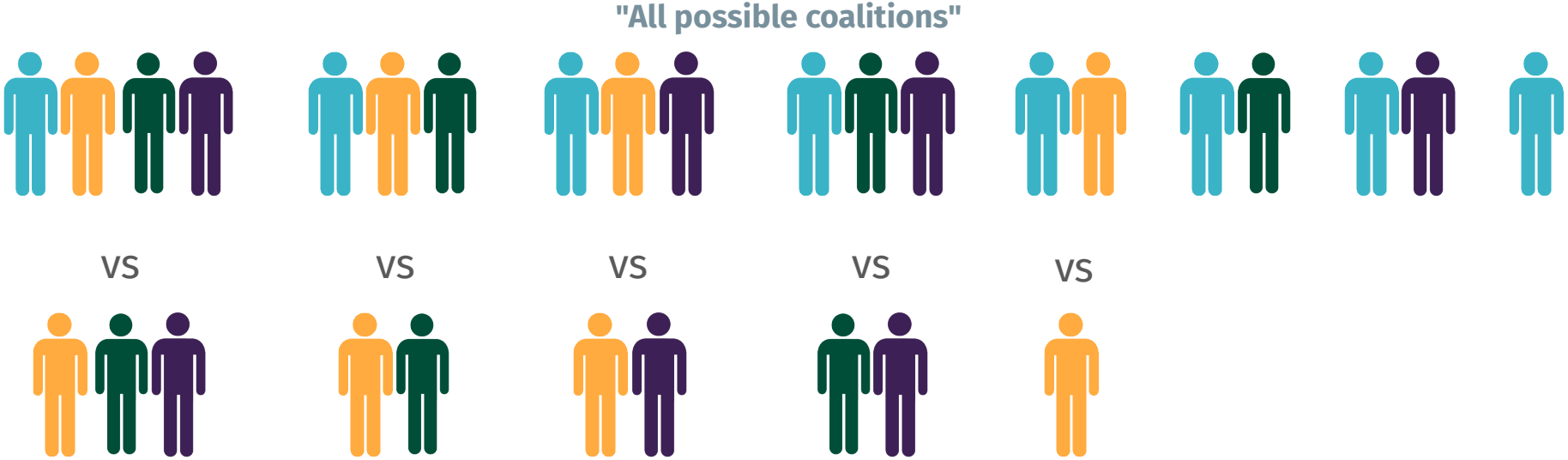
- **Shapley Values:** a method from coalitional game theory, which measures the *average marginal contribution* of a feature value across all possible *coalitions*
- **Idea:** Suppose we wanted to measure the importance of the blue person

"All possible coalitions"



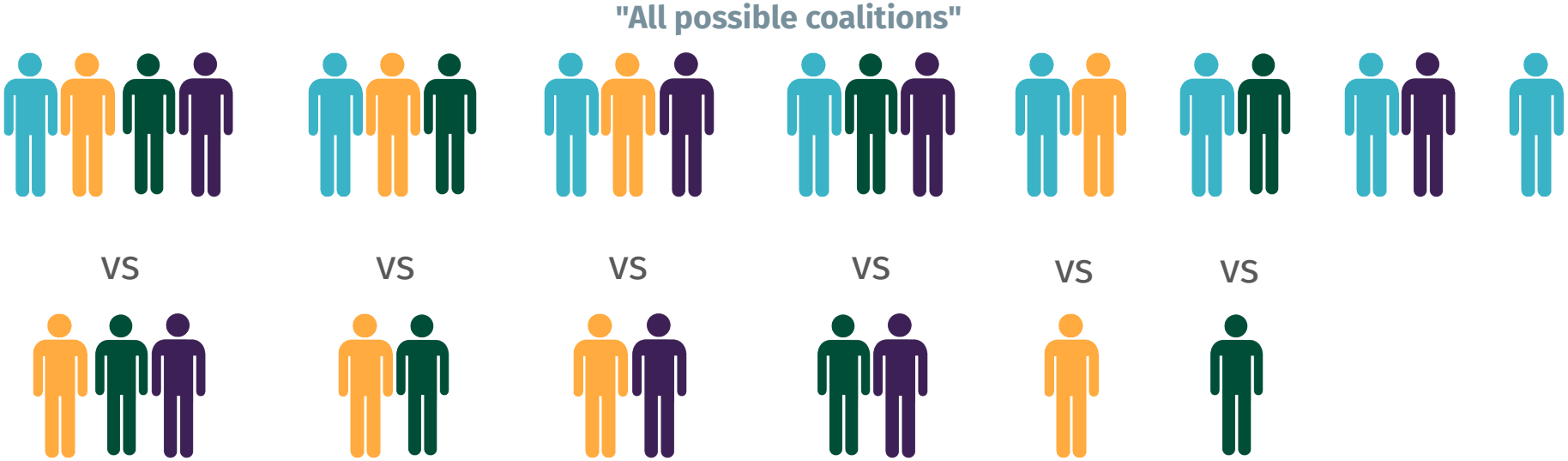
Shapley Values in Game Theory

- **Shapley Values:** a method from coalitional game theory, which measures the *average marginal contribution* of a feature value across all possible *coalitions*
- **Idea:** Suppose we wanted to measure the importance of the blue person



Shapley Values in Game Theory

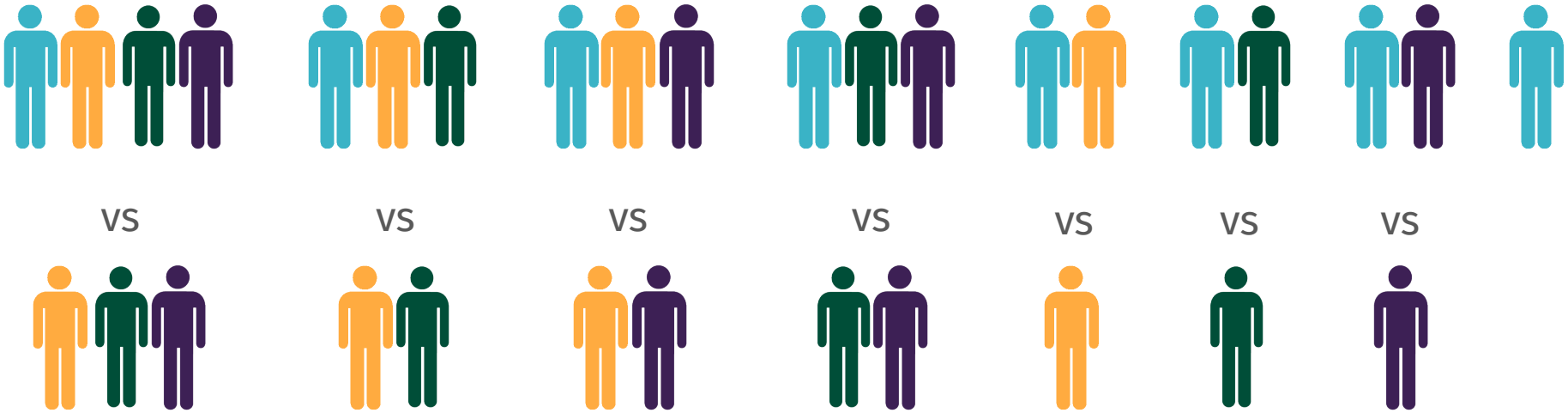
- **Shapley Values:** a method from coalitional game theory, which measures the *average marginal contribution* of a feature value across all possible *coalitions*
- **Idea:** Suppose we wanted to measure the importance of the blue person



Shapley Values in Game Theory

- **Shapley Values:** a method from coalitional game theory, which measures the *average marginal contribution* of a feature value across all possible *coalitions*
- **Idea:** Suppose we wanted to measure the importance of the blue person

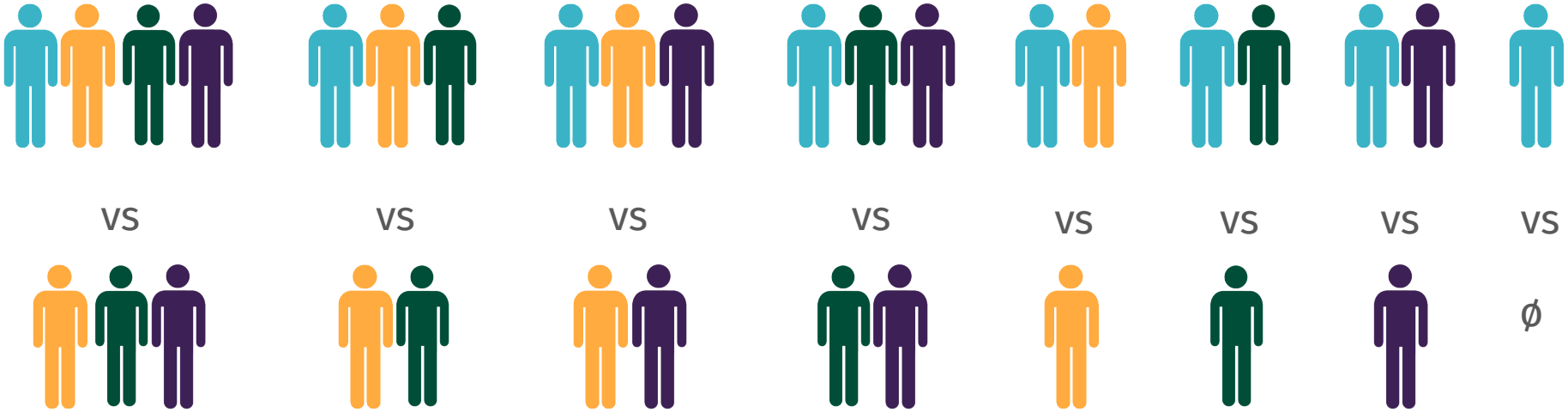
"All possible coalitions"



Shapley Values in Game Theory

- **Shapley Values:** a method from coalitional game theory, which measures the *average marginal contribution* of a feature value across all possible *coalitions*
- **Idea:** Suppose we wanted to measure the importance of the blue person

"All possible coalitions"



Shapley Values in Game Theory

- **Shapley Values:** a method from coalitional game theory, which measures the *average marginal contribution* of a feature value across all possible *coalitions*
- **Idea:** Suppose we wanted to measure the importance of the blue person

"All possible coalitions"

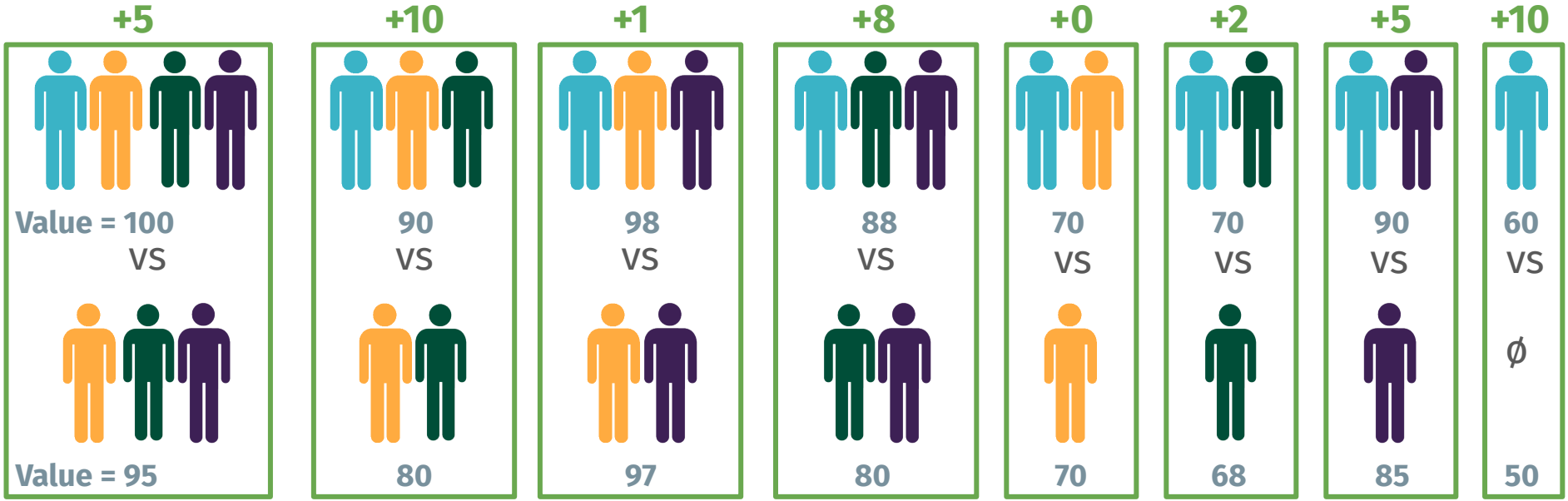
Shapley Values in Game Theory

- **Shapley Values:** a method from coalitional game theory, which measures the *average marginal contribution* of a feature value across all possible *coalitions*
- **Idea:** Suppose we wanted to measure the importance of the blue person

"All possible coalitions"

Shapley Values in Game Theory

- **Shapley Values:** a method from coalitional game theory, which measures the *average marginal contribution* of a feature value across all possible *coalitions*
- **Idea:** Suppose we wanted to measure the importance of the blue person



Shapley Values in Game Theory

- **Shapley Values:** a method from coalitional game theory, which measures the *average marginal contribution* of a feature value across all possible *coalitions*
- In the example, the Shapley value for **player** *j* is:

$$\phi_j(val) = \sum_{S \subseteq \{1,\ldots,p\} \setminus \{j\}} \frac{|S|!\,(p - |S| - 1)!}{p!} \left(val\left(S \cup \{j\}\right) - val(S)\right)$$

Shapley Values in Game Theory

- **Shapley Values:** a method from coalitional game theory, which measures the *average marginal contribution* of a feature value across all possible *coalitions*
- In the example, the Shapley value for **player** *j* is:

$$\phi_j(val) = \sum_{S \subseteq \{1,\ldots,p\} \setminus \{j\}} \frac{|S|!\,(p - |S| - 1)!}{p!} \left(val\left(S \cup \{j\}\right) - val(S)\right)$$

coalition

Shapley Values in Game Theory

- **Shapley Values:** a method from coalitional game theory, which measures the *average marginal contribution* of a feature value across all possible *coalitions*
- In the example, the Shapley value for **player** *j* is:

$$\phi_j(val) = \sum_{S \subseteq \{1,\ldots,p\} \setminus \{j\}} \frac{|S|!\,(p - |S| - 1)!}{p!} \left(val\left(S \cup \{j\}\right) - val(S)\right)$$

coalition

Average over all coalitions

Shapley Values in Game Theory

- **Shapley Values:** a method from coalitional game theory, which measures the *average marginal contribution* of a feature value across all possible *coalitions*
- In the example, the Shapley value for **player** *j* is:

$$\phi_j(val) = \underbrace{\sum_{S \subseteq \{1,\ldots,p\} \setminus \{j\}}}_{\text{Average over all coalitions}} \underbrace{\frac{|S|!\,(p - |S| - 1)!}{p!}}_{\text{coalition}} \underbrace{(val\,(S \cup \{j\}) - val(S))}_{\text{Value with person } j}$$

Shapley Values in Game Theory

- **Shapley Values:** a method from coalitional game theory, which measures the *average marginal contribution* of a feature value across all possible *coalitions*
- In the example, the Shapley value for **player** *j* is:

$$\phi_j(val) = \sum_{S \subseteq \{1,\ldots,p\} \setminus \{j\}} \frac{|S|!\,(p - |S| - 1)!}{p!} \underbrace{(val\,(S \cup \{j\})}_{\text{Value with person } j} - \underbrace{val(S))}_{\text{Value without person } j}$$

coalition

Average over all coalitions

Shapley Values in Game Theory

- **Shapley Values:** a method from coalitional game theory, which measures the *average marginal contribution* of a feature value across all possible *coalitions*
- In the example, the Shapley value for **player** *j* is:

$$\phi_j(val) = \sum_{S \subseteq \{1,\ldots,p\} \setminus \{j\}} \frac{|S|!\,(p - |S| - 1)!}{p!} \left(val\left(S \cup \{j\}\right) - val(S)\right)$$

- coalition
- Average over all coalitions
- Normalization weight*
- Value **with** person *j*
- Value **without** person *j*

Shapley Values in Game Theory

- **Shapley Values:** a method from coalitional game theory, which measures the *average marginal contribution* of a feature value across all possible *coalitions*
- In the example, the Shapley value for **player** *j* is:

$$\phi_j(val) = \underbrace{\sum_{S \subseteq \{1,\ldots,p\} \setminus \{j\}}}_{\text{Average over all coalitions}} \underbrace{\frac{|S|!\,(p-|S|-1)!}{p!}}_{\text{Normalization weight*}} \underbrace{(val\,(S \cup \{j\})}_{\text{Value with person } j} - \underbrace{val(S))}_{\text{Value without person } j}$$

coalition

where *val* is the "value" function (the exam score in the example)

Shapley Values for an ML prediction model

- **Shapley Values:** a method from coalitional game theory, which measures the *average marginal contribution* of a feature value across all possible *coalitions*
- **For an ML model *f***, the (original) Shapley value for **feature** *j* is:

$$\phi_j(val) = \sum_{S \subseteq \{1,\ldots,p\} \setminus \{j\}} \underbrace{\frac{|S|!\,(p-|S|-1)!}{p!}}_{\text{Normalization weight*}} \underbrace{(val\,(S \cup \{j\}) - val(S))}_{\text{Value with person } j \text{ minus Value without person } j}$$

coalition

Average over all coalitions

where *val*(*S*) is the **prediction error** from the model *f* trained using the subset of features in *S*

Shapley Values for an ML prediction model

- **Shapley Values:** a method from coalitional game theory, which measures the *average marginal contribution* of a feature value across all possible *coalitions*
- **For an ML model *f*,** the (original) Shapley value for **feature** *j* is:

$$\phi_j(val) = \sum_{S \subseteq \{1,\ldots,p\} \setminus \{j\}} \frac{|S|!\,(p-|S|-1)!}{p!} \left(val\left(S \cup \{j\}\right) - val(S)\right)$$

- coalition
- $S \subseteq \{1,\ldots,p\} \setminus \{j\}$: Average over all coalitions
- $\frac{|S|!\,(p-|S|-1)!}{p!}$: Normalization weight*
- $val\left(S \cup \{j\}\right)$: Value **with** person *j*
- $val(S)$: Value **without** person *j*

where *val*(*S*) is the **prediction error** from the model *f* trained using the subset of features in *S*

Problem: infeasible to re-fit *f* for every possible coalition of features

Shapley Values for an ML prediction model without re-fitting

- **Shapley Values:** a method from coalitional game theory, which measures the *average marginal contribution* of a feature value across all possible *coalitions*
- **For an ML model *f***, the (local) Shapley value for **feature** *j* and **sample** *x* is:

$$\phi_j(val) = \sum_{S \subseteq \{1,\ldots,p\} \setminus \{j\}} \underbrace{\frac{|S|!\,(p-|S|-1)!}{p!}}_{\text{Normalization weight*}} \underbrace{(val\,(S \cup \{j\}) - val(S))}_{\text{Value with person } j \text{ minus value without person } j}$$

coalition

Average over all coalitions

Shapley Values for an ML prediction model without re-fitting

- **Shapley Values:** a method from coalitional game theory, which measures the *average marginal contribution* of a feature value across all possible *coalitions*
- **For an ML model *f***, the (local) Shapley value for **feature** *j* and **sample** *x* is:

$$\phi_j(val) = \sum_{S \subseteq \{1,\ldots,p\} \setminus \{j\}} \frac{|S|!\,(p-|S|-1)!}{p!} \left(val\left(S \cup \{j\}\right) - val(S)\right)$$

- coalition
- Average over all coalitions
- Normalization weight*
- Value **with** person *j*
- Value **without** person *j*

where $val(S) = \mathbb{E}[\hat{f}(X) \mid X_S = x_S] - \mathbb{E}[\hat{f}(X)]$

- Average prediction, conditioned on observing x_S
- Average prediction

Properties of Shapley Values

Shapley values are the only solution that satisfies the four properties below:

Properties of Shapley Values

Shapley values are the only solution that satisfies the four properties below:

- **Efficiency:** The feature contributions must add up to the difference between the prediction for x and the average

$$\sum_{j=1}^{p} \phi_j = \hat{f}(x) - E_X(\hat{f}(X))$$

Properties of Shapley Values

Shapley values are the only solution that satisfies the four properties below:

- **Efficiency:** The feature contributions must add up to the difference between the prediction for x and the average

$$\sum_{j=1}^{p} \phi_j = \hat{f}(x) - E_X(\hat{f}(X))$$

- **Symmetry:** The contributions of two features *j* and *k* are equal if they contribute equally to all possible coalitions
  - If $val(S \cup \{j\}) = val(S \cup \{k\})$ for all $S \subseteq \{1, \ldots, p\} \backslash \{j, k\}$, then $\phi_j = \phi_k$

Properties of Shapley Values

Shapley values are the only solution that satisfies the four properties below:

- **Efficiency:** The feature contributions must add up to the difference between the prediction for x and the average

$$\sum_{j=1}^{p} \phi_j = \hat{f}(x) - E_X(\hat{f}(X))$$

- **Symmetry:** The contributions of two features *j* and *k* are equal if they contribute equally to all possible coalitions
  - If $val(S \cup \{j\}) = val(S \cup \{k\})$ for all $S \subseteq \{1, \ldots, p\} \backslash \{j, k\}$, then $\phi_j = \phi_k$

- **Dummy:** A feature *j* that does not change the predicted value (no matter the choice of coalition) should have a contribution of 0
  - If $val(S \cup \{j\}) = val(S)$ for all $S \subseteq \{1, \ldots, p\}$, then $\phi_j = 0$

Properties of Shapley Values

Shapley values are the only solution that satisfies the four properties below:

- **Efficiency:** The feature contributions must add up to the difference between the prediction for x and the average

$$\sum_{j=1}^{p} \phi_j = \hat{f}(x) - E_X(\hat{f}(X))$$

- **Symmetry:** The contributions of two features *j* and *k* are equal if they contribute equally to all possible coalitions
  - If $val(S \cup \{j\}) = val(S \cup \{k\})$ for all $S \subseteq \{1, \ldots, p\} \backslash \{j, k\}$, then $\phi_j = \phi_k$

- **Dummy:** A feature *j* that does not change the predicted value (no matter the choice of coalition) should have a contribution of 0
  - If $val(S \cup \{j\}) = val(S)$ for all $S \subseteq \{1, \ldots, p\}$, then $\phi_j = 0$

- **Additivity:** For a game with additive payouts, the Shapley values are also additive
  - For a RF, the Shapley value of the RF = average of the Shapley values of each tree

How do we estimate Shapley values?

- We can try to plug in sample estimators for the expectations

$$\phi_j(val) = \underbrace{\sum_{S \subseteq \{1,\ldots,p\} \setminus \{j\}}}_{\text{Average over all coalitions}} \underbrace{\frac{|S|!\,(p - |S| - 1)!}{p!}}_{\text{Normalization weight}} \underbrace{(val\,(S \cup \{j\})}_{\text{Value with feature } j} - \underbrace{val(S))}_{\text{Value without feature } j}$$

coalition

where $val(S) = \underbrace{\mathbb{E}[\hat{f}(X) \mid X_S = x_S]}_{\text{Average prediction, conditioned on observing } x_S} - \underbrace{\mathbb{E}[\hat{f}(X)]}_{\text{Average prediction}}$

How do we estimate Shapley values?

+ We can try to plug in sample estimators for the expectations

$$\phi_j(val) = \sum_{S \subseteq \{1,\ldots,p\} \setminus \{j\}} \underbrace{\frac{|S|!\,(p - |S| - 1)!}{p!}}_{\text{Normalization weight}} \underbrace{(val\,(S \cup \{j\}) - val(S))}_{\text{Value with feature } j \text{ minus value without feature } j}$$

coalition

Average over all coalitions

where $val(S) = \underbrace{\mathbb{E}[\hat{f}(X) \mid X_S = x_S]}_{\text{Average prediction, conditioned on observing } x_S} - \underbrace{\mathbb{E}[\hat{f}(X)]}_{\text{Average prediction}}$

e.g., if $S = \{1, 2, 5\}$:

Randomly sampled values from features 3, 4, and 6

How do we estimate Shapley values?

+ We can try to plug in sample estimators for the expectations

$$\phi_j(val) = \sum_{S \subseteq \{1,\ldots,p\} \setminus \{j\}} \frac{|S|!\,(p - |S| - 1)!}{p!} \left(val\left(S \cup \{j\}\right) - val(S)\right)$$

- coalition
- $S \subseteq \{1,\ldots,p\} \setminus \{j\}$: Average over all coalitions
- $\frac{|S|!\,(p - |S| - 1)!}{p!}$: Normalization weight
- $val\left(S \cup \{j\}\right)$: Value **with** feature *j*
- $val(S)$: Value **without** feature *j*

where $val(S) = \underbrace{\mathbb{E}[\hat{f}(X) \mid X_S = x_S]}_{\text{Average prediction, conditioned on observing } x_S} - \underbrace{\mathbb{E}[\hat{f}(X)]}_{\text{Average prediction}}$

Problem: still computationally-expensive to estimate for every possible coalition of features

SHAP (SHapley Additive exPlanations)

SHAP (SHapley Additive exPlanations)

- **SHAP:** computationally-feasible approach to approximate Shapley values via a surrogate explanation model g

SHAP (SHapley Additive exPlanations)

- **SHAP:** computationally-feasible approach to approximate Shapley values via a surrogate explanation model *g*
  - Aims to approximate a prediction model via an explanation model *g*:

SHAP (SHapley Additive exPlanations)

- **SHAP:** computationally-feasible approach to approximate Shapley values via a surrogate explanation model *g*
  - Aims to approximate a prediction model via an explanation model *g*:

$$g_i(z') = \phi_0 + \sum_{j=1}^{p} \phi_j z'_j$$

where $z' \in \{0,1\}^p$ is a coalition vector

ϕ_j is the contribution of feature *j*

SHAP (SHapley Additive exPlanations)

- **SHAP:** computationally-feasible approach to approximate Shapley values via a surrogate explanation model *g*
  - Aims to approximate a prediction model via an explanation model *g*:

$$g_i(z') = \phi_0 + \sum_{j=1}^{p} \phi_j z'_j$$

where $z' \in \{0,1\}^p$ is a coalition vector

ϕ_j is the contribution of feature *j*

- Different forms of SHAP:

SHAP (SHapley Additive exPlanations)

- **SHAP:** computationally-feasible approach to approximate Shapley values via a surrogate explanation model *g*
  - Aims to approximate a prediction model via an explanation model *g*:

$$g_i(z') = \phi_0 + \sum_{j=1}^{p} \phi_j z'_j$$

where $z' \in \{0,1\}^p$ is a coalition vector

ϕ_j is the contribution of feature *j*

- Different forms of SHAP:
  - **KernelSHAP:** for general models

SHAP (SHapley Additive exPlanations)

- **SHAP:** computationally-feasible approach to approximate Shapley values via a surrogate explanation model *g*
  - Aims to approximate a prediction model via an explanation model *g*:

$$g_i(z') = \phi_0 + \sum_{j=1}^{p} \phi_j z'_j$$

where $z' \in \{0,1\}^p$ is a coalition vector

ϕ_j is the contribution of feature *j*

- Different forms of SHAP:
  - **KernelSHAP:** for general models
  - **TreeSHAP:** for tree-based models (exploits tree structure for more efficient computation)

SHAP (SHapley Additive exPlanations)

- **SHAP:** computationally-feasible approach to approximate Shapley values via a surrogate explanation model *g*
  - Aims to approximate a prediction model via an explanation model *g*:

$$g_i(z') = \phi_0 + \sum_{j=1}^{p} \phi_j z'_j$$

where $z' \in \{0,1\}^p$ is a coalition vector

ϕ_j is the contribution of feature *j*

- Different forms of SHAP:
  - **KernelSHAP:** for general models
  - **TreeSHAP:** for tree-based models (exploits tree structure for more efficient computation)
  - **DeepSHAP:** for neural networks exploits backpropagation for more efficient computation)

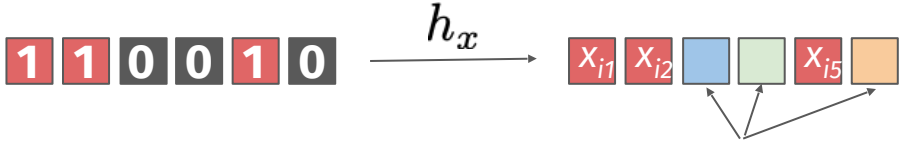
KernelSHAP

KernelSHAP

1. Sample coalitions $z'_k \in \{0,1\}^p$ 1 1 0 0 1 0
   (1 = feature in coalition; 0 = feature not in coalition)

KernelSHAP

1. Sample coalitions $z'_k \in \{0,1\}^p$ 1 1 0 0 1 0
   (1 = feature in coalition; 0 = feature not in coalition)
2. Let $h_x(z'_k)$ denote the data vector that is equal to *x* for features in the coalition and take subsampled feature values for features not in the coalition

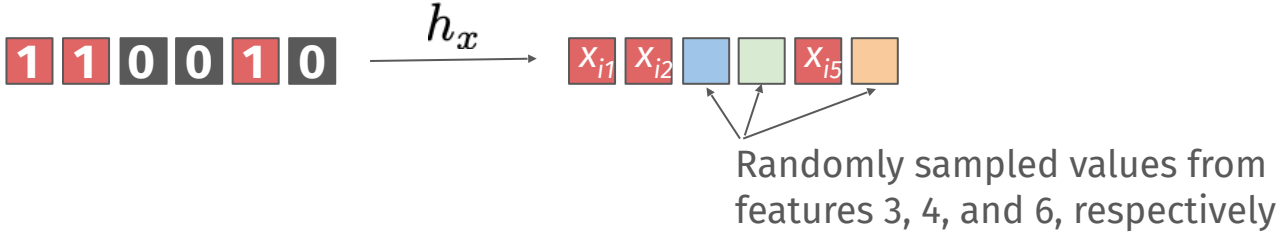


Randomly sampled values from features 3, 4, and 6, respectively

KernelSHAP

1. Sample coalitions $z'_k \in \{0,1\}^p$ 

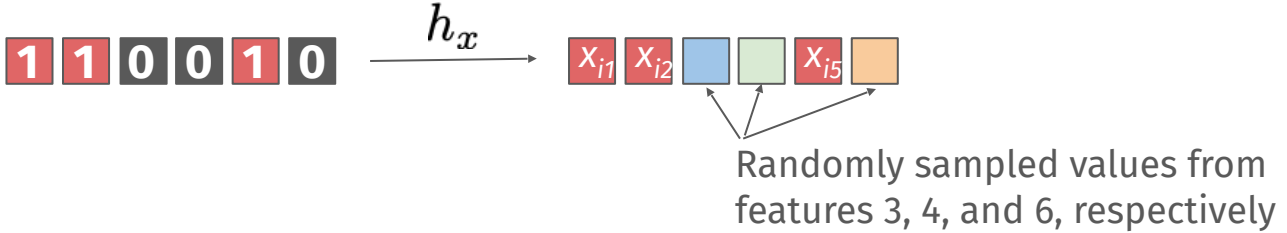
(1 = feature in coalition; 0 = feature not in coalition)
2. Let $h_x(z'_k)$ denote the data vector that is equal to *x* for features in the coalition and take subsampled feature values for features not in the coalition



3. Get prediction for each $h_x(z'_k) \longrightarrow \hat{f}(h_x(z'_k))$

KernelSHAP

1. Sample coalitions $z'_k \in \{0,1\}^p$ **1 1 0 0 1 0**
   (1 = feature in coalition; 0 = feature not in coalition)
2. Let $h_x(z'_k)$ denote the data vector that is equal to *x* for features in the coalition and take subsampled feature values for features not in the coalition

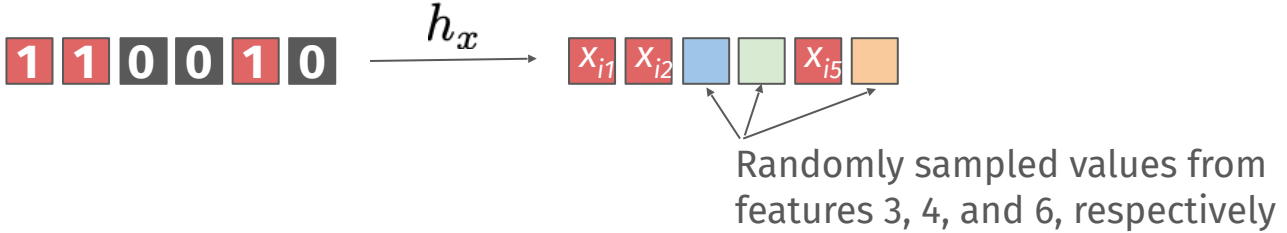


3. Get prediction for each $h_x(z'_k) \longrightarrow \hat{f}(h_x(z'_k))$
4. Fit weighted linear model (same ideas as LIME):

$$L(\hat{f}, g, \pi_x) = \sum_{z' \in Z} [\hat{f}(h_x(z')) - g(z')]^2 \pi_x(z') \quad \text{where} \quad \pi_x(z') = \frac{(M-1)}{\binom{M}{|z'|}|z'|(M-|z'|)}$$

KernelSHAP

1. Sample coalitions $z'_k \in \{0,1\}^p$ **1 1 0 0 1 0**
   (1 = feature in coalition; 0 = feature not in coalition)
2. Let $h_x(z'_k)$ denote the data vector that is equal to *x* for features in the coalition and take subsampled feature values for features not in the coalition



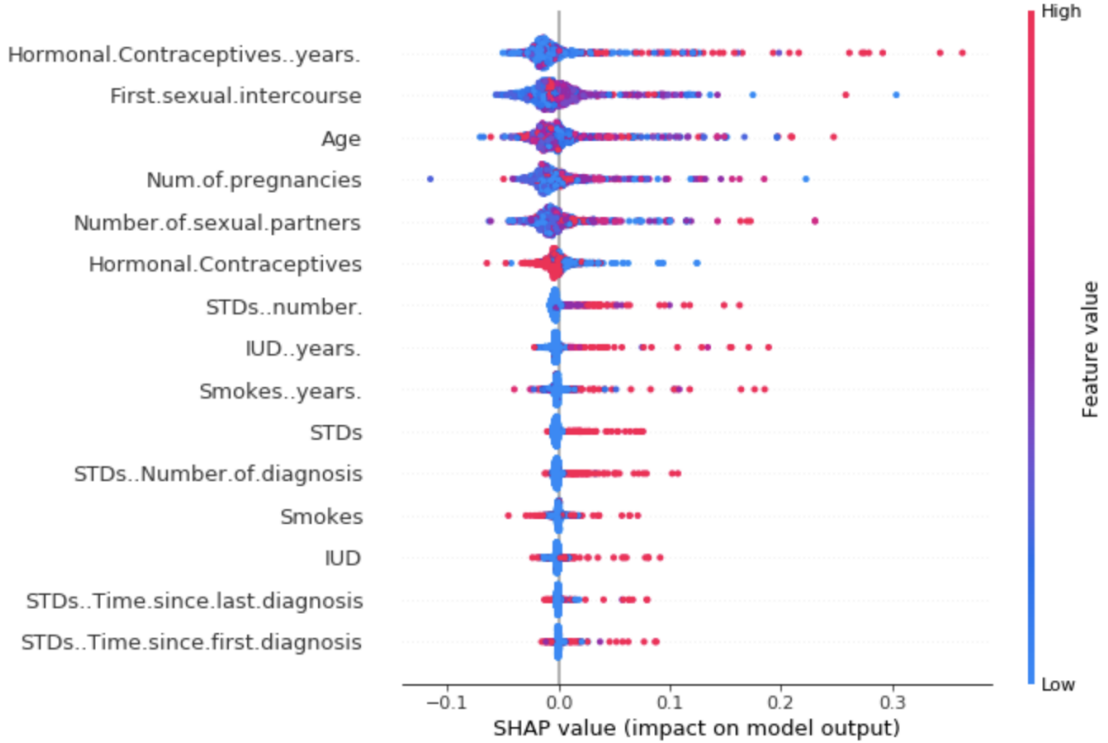
3. Get prediction for each $h_x(z'_k) \longrightarrow \hat{f}(h_x(z'_k))$
4. Fit weighted linear model (same ideas as LIME):

$$L(\hat{f}, g, \pi_x) = \sum_{z' \in Z} [\hat{f}(h_x(z')) - g(z')]^2 \pi_x(z') \quad \text{where} \quad \pi_x(z') = \frac{(M-1)}{\binom{M}{|z'|}|z'|(M-|z'|)}$$

Estimated coefficients correspond to the SHAP local feature contributions

Visualizations with SHAP/KernelSHAP

Predicting cancer risk



SHAP: Practical Considerations

- SHAP is THE most popular local feature importance method
- Sampling destroys natural correlation structure
  - This sampling may result in problematic extrapolations
  - More on issues with correlated features next time
- Like LIME, SHAP is interpreting a particular **fixed** model
  - No refitting involved
  - Should NOT be interpreted as the feature importance if we had removed the feature and retrained the model
- Can also be unstable and manipulated by adversarial attacks to hide biases*
- Shapley and SHAP are NOT the same thing

* Slack et al. (2020) Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods.