Interpretable Machine Learning

March 31, 2026

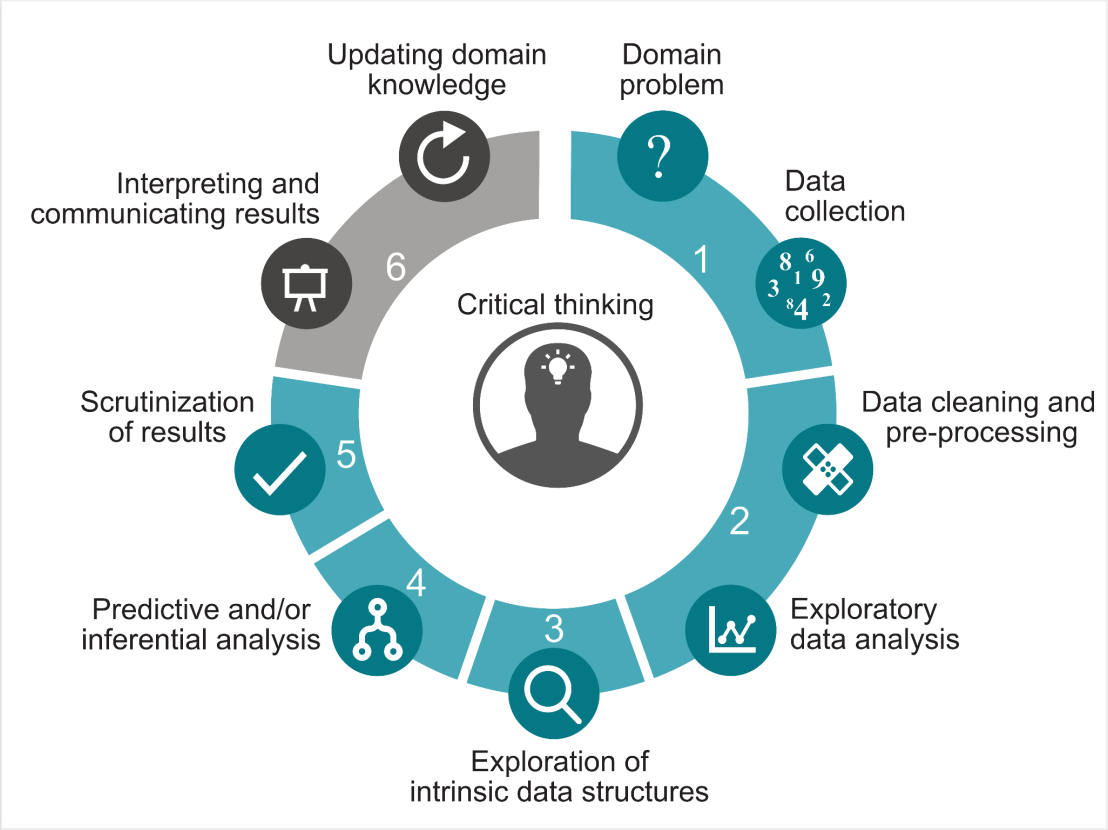
Announcements

- All done with labs!
- **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**

Looking ahead to the rest of the semester

- This week: interpretability/explainability
- Next Tuesday: code review with GitHub flow
- Next four lectures: large language models (LLMs)
- Last two classes: in-class final project presentations (~6-8 min each)

The Big Picture: Data Science Life Cycle



Credits: Veridical Data Science (VDS) Textbook

Plan for this week: Interpretability/Explainability

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

Why do we care?


MARVEL
AVENGERS
AGE OF ULTRON


TOM CRUISE
MISSION: IMPOSSIBLE
DEAD RECKONING
PART ONE
ONLY IN THEATRES
JULY 12

Our world is increasingly being run by algorithms

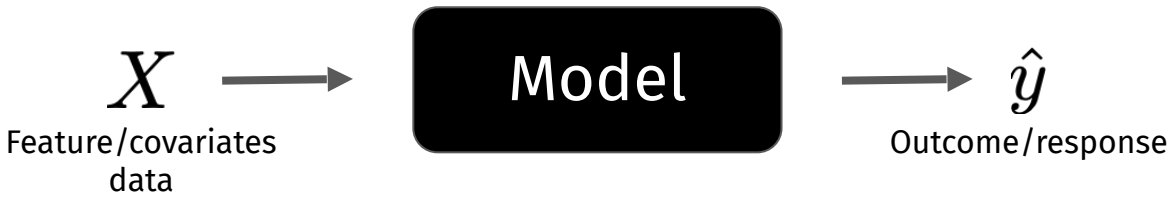
Algorithms are being used to make high-stakes decisions on

- Whom to give loans to
- Whom to give job interviews to
- Whom to grant parole to
- ...

Our world is increasingly being run by algorithms

Algorithms are being used to make high-stakes decisions on

- Whom to give loans to
- Whom to give job interviews to
- Whom to grant parole to
- ...



If our algorithms are “black-boxes”…

Glenn Rodriguez

The New York Times

When a Computer Program Keeps You in Jail

By Rebecca Wexler

June 13, 2017

Sally Deng

The criminal justice system is becoming automated. At every stage — from policing and investigations to bail, evidence, sentencing and parole — computer systems play a role. Artificial intelligence deploys cops on the beat. Audio sensors generate gunshot alerts.

Source: NY Times

If our algorithms are “black-boxes”...

Glenn Rodriguez

Take the case of Glenn Rodríguez. An inmate at the Eastern Correctional Facility in upstate New York, Mr. Rodríguez was denied parole last year despite having a nearly perfect record of rehabilitation. The reason? A high score from a computer system called Compas. The company that makes Compas considers the weighting of inputs to be proprietary information. That forced Mr. Rodríguez to rely on his own ingenuity to figure out what had gone wrong.

Source: NY Times

If our algorithms are “black-boxes”...

Glenn Rodriguez

Take the case of Glenn Rodríguez. An inmate at the Eastern Correctional Facility in upstate New York, Mr. Rodríguez was denied parole last year despite having a nearly perfect record of rehabilitation. The reason? A high score from a computer system called Compas. The company that makes Compas considers the weighting of inputs to be proprietary information. That forced Mr. Rodríguez to rely on his own ingenuity to figure out what had gone wrong.

This year, Mr. Rodríguez returned to the parole board with the same faulty Compas score. He had identified an error in one of the inputs for his Compas assessment. But without knowing the input weights, he was unable to explain the effect of this error, or persuade anyone to correct it. Instead of challenging the result, he was left to try to argue for parole despite the result.

Source: NY Times

Laws and Regulations on Algorithmic Transparency

Laws and Regulations on Algorithmic Transparency

- **European Union General Data Protection Regulation** (GDPR) (2018)
  - *"Right to Explanation"*: entities must "ensure fair and transparent processing" of personal data and provide citizens with access to "meaningful information about the logic involved" in certain automated decision-making systems

Laws and Regulations on Algorithmic Transparency

- **European Union General Data Protection Regulation** (GDPR) (2018)
  - *"Right to Explanation"*: entities must "ensure fair and transparent processing" of personal data and provide citizens with access to "meaningful information about the logic involved" in certain automated decision-making systems
- **European Union AI Act** (in effect since August 1, 2024) [summary]
  - Prohibits some AI uses outright and implements strict governance, risk management and transparency requirements for others

Laws and Regulations on Algorithmic Transparency

- **European Union General Data Protection Regulation** (GDPR) (2018)
  - *"Right to Explanation"*: entities must "ensure fair and transparent processing" of personal data and provide citizens with access to "meaningful information about the logic involved" in certain automated decision-making systems
- **European Union AI Act** (in effect since August 1, 2024) [summary]
  - Prohibits some AI uses outright and implements strict governance, risk management and transparency requirements for others
- **US Dodd-Frank Act** (2010)
  - Requires that creditors disclose certain information when making credit decisions, e.g., providing key factors that negatively affected the credit score when a person is denied credit

Laws and Regulations on Algorithmic Transparency

- **European Union General Data Protection Regulation** (GDPR) (2018)
  - *"Right to Explanation"*: entities must "ensure fair and transparent processing" of personal data and provide citizens with access to "meaningful information about the logic involved" in certain automated decision-making systems
- **European Union AI Act** (in effect since August 1, 2024) [summary]
  - Prohibits some AI uses outright and implements strict governance, risk management and transparency requirements for others
- **US Dodd-Frank Act** (2010)
  - Requires that creditors disclose certain information when making credit decisions, e.g., providing key factors that negatively affected the credit score when a person is denied credit
- **Many others to come...**

Interpretability in machine learning is increasingly necessary

Interpretability in machine learning is increasingly necessary

+ Answers the “how” (important for understanding the *science/mechanism*)

Interpretability in machine learning is increasingly necessary

- Answers the “how” (important for understanding the *science/mechanism*)
- Instills trust/distrust in a model

Interpretability in machine learning is increasingly necessary

- Answers the “how” (important for understanding the *science/mechanism*)
- Instills trust/distrust in a model
- Aids human-in-the-loop workflow
  [can be used to iterate and improve the analysis/modeling process]

Interpretability in machine learning is increasingly necessary

- Answers the “how” (important for understanding the *science/mechanism*)
- Instills trust/distrust in a model
- Aids human-in-the-loop workflow
  [can be used to iterate and improve the analysis/modeling process]
- Facilitates auditing for errors and biases

Interpretability in machine learning is increasingly necessary

- Answers the “how” (important for understanding the *science/mechanism*)
- Instills trust/distrust in a model
- Aids human-in-the-loop workflow
  [can be used to iterate and improve the analysis/modeling process]
- Facilitates auditing for errors and biases

However, this is NOT a complete solution to eliminating biases in machine learning

Interpretability in machine learning is increasingly necessary

- Answers the "how" (important for understanding the *science/mechanism*)
- Instills trust/distrust in a model
- Aids human-in-the-loop workflow
  [can be used to iterate and improve the analysis/modeling process]
- Facilitates auditing for errors and biases

However, this is NOT a complete solution to eliminating biases in machine learning

We also need: post-hoc EDA and "local" investigations, careful curation and critique of the input data, transparent documentation of analytical process, …

What do we mean by “interpretable”?

Which models are “interpretable" or “not interpretable”?

Which models are “interpretable" or “not interpretable”?

The meaning of "interpretable" is subjective and depends on your context and audience!

How can we interpret our machine learning models?

Avenue 1: “Let’s develop interpretable* models from the start”

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

How can we interpret our machine learning models?

Avenue 1: “Let’s develop interpretable* models from the start”

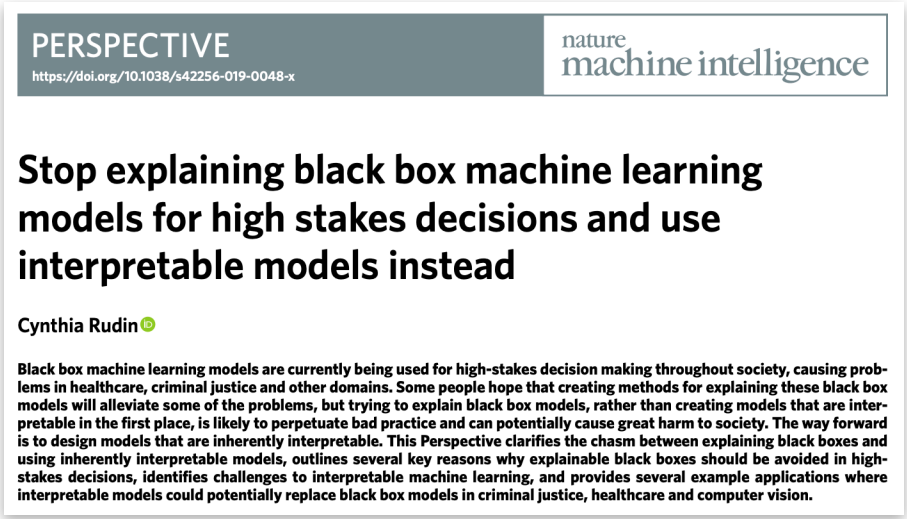
Can be visualized.
Implemented by hand.
Mirrors way people think.

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

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.



PERSPECTIVE
https://doi.org/10.1038/s42256-019-0048-x

nature machine intelligence

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

Cynthia Rudin

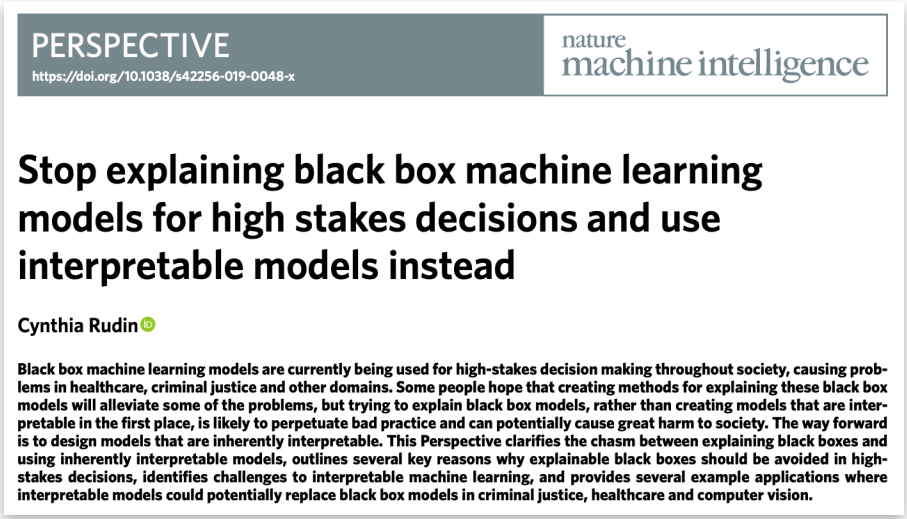
Black box machine learning models are currently being used for high-stakes decision making throughout society, causing problems in healthcare, criminal justice and other domains. Some people hope that creating methods for explaining these black box models will alleviate some of the problems, but trying to explain black box models, rather than creating models that are interpretable in the first place, is likely to perpetuate bad practice and can potentially cause great harm to society. The way forward is to design models that are inherently interpretable. This Perspective clarifies the chasm between explaining black boxes and using inherently interpretable models, outlines several key reasons why explainable black boxes should be avoided in high-stakes decisions, identifies challenges to interpretable machine learning, and provides several example applications where interpretable models could potentially replace black box models in criminal justice, healthcare and computer vision.

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

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.



PERSPECTIVE
https://doi.org/10.1038/s42256-019-0048-x

nature machine intelligence

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

Cynthia Rudin

Black box machine learning models are currently being used for high-stakes decision making throughout society, causing problems in healthcare, criminal justice and other domains. Some people hope that creating methods for explaining these black box models will alleviate some of the problems, but trying to explain black box models, rather than creating models that are interpretable in the first place, is likely to perpetuate bad practice and can potentially cause great harm to society. The way forward is to design models that are inherently interpretable. This Perspective clarifies the chasm between explaining black boxes and using inherently interpretable models, outlines several key reasons why explainable black boxes should be avoided in high-stakes decisions, identifies challenges to interpretable machine learning, and provides several example applications where interpretable models could potentially replace black box models in criminal justice, healthcare and computer vision.

- In many cases, it's possible to learn an interpretable model that gives just as good prediction accuracy as a black-box model (e.g., neural network)
- It just takes more time and critical thinking to find this interpretable model

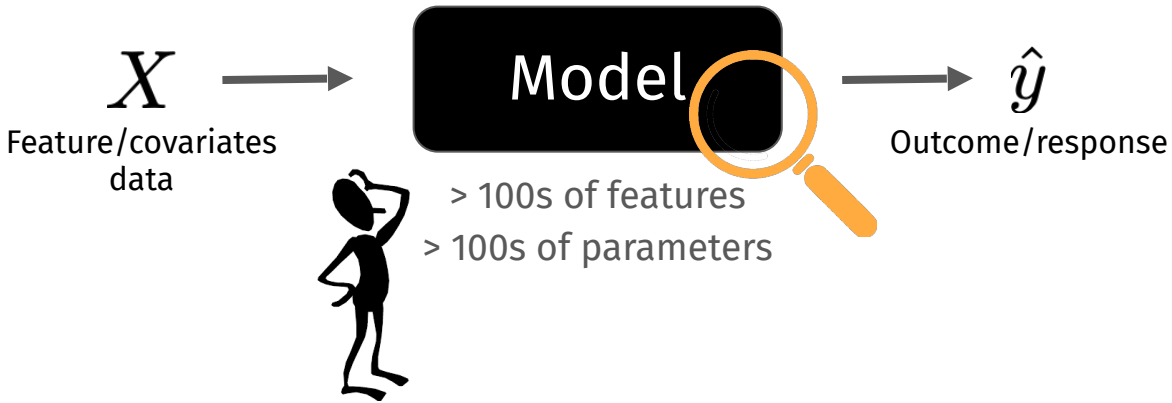
* Many technical definitions… (e.g., Murdoch et al. (2019))

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.

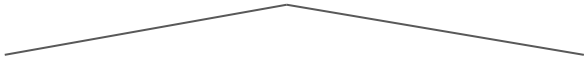
Avenue 2: "Let's develop tools to interpret complex, black-box models"



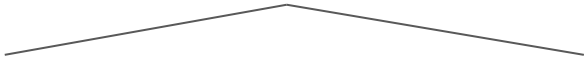
* Many technical definitions... (e.g., Murdoch et al. (2019))

Intro to Interpretability Tools

(One possible) Taxonomy of interpretability tools



(One possible) Taxonomy of interpretability tools



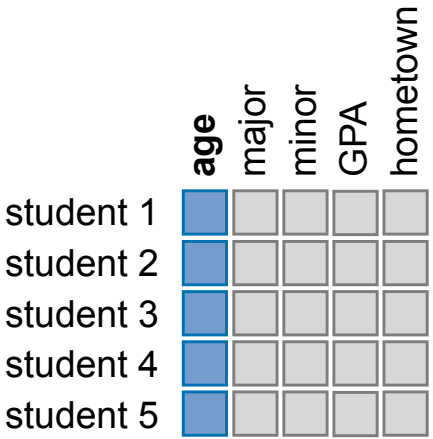
Feature Importances

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

(One possible) Taxonomy of interpretability tools

Feature Importances

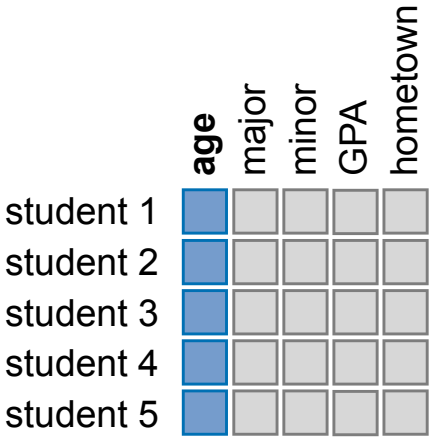
How important is a **feature** (or set of features) in making the model predictions?



(One possible) Taxonomy of interpretability tools

Feature Importances

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



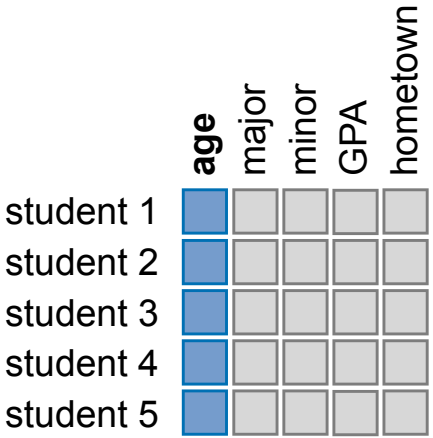
Sample Influences

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

(One possible) Taxonomy of interpretability tools

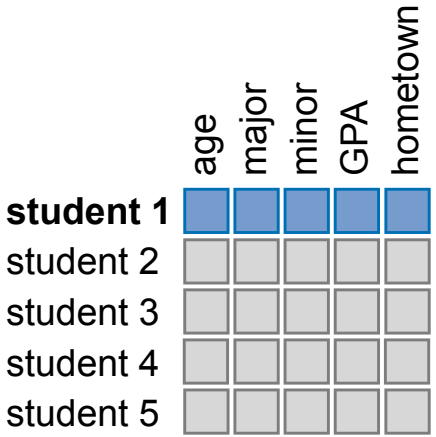
Feature Importances

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



Sample Influences

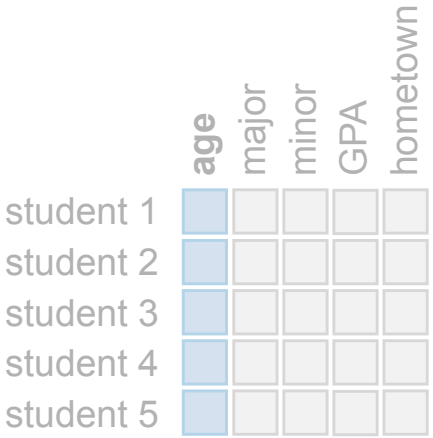
How does leaving out a **sample** change the model fit, predictions, and/or error?



(One possible) Taxonomy of interpretability tools

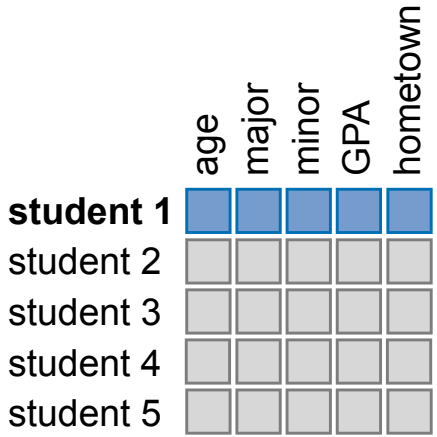
Feature Importances

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?



Sample Influences

Sample Influence

Sample Influence

Q: How does a sample affect the prediction model fit? In particular, what would have happened if we had left out that sample from the model training?

Sample Influence

Q: How does a sample affect the prediction model fit? In particular, what would have happened if we had left out that sample from the model training?

Computationally expensive to leave-one sample out and refit model for even moderate-sized datasets

Sample Influence

Q: How does a sample affect the prediction model fit? In particular, what would have happened if we had left out that sample from the model training?

Computationally expensive to leave-one sample out and refit model for even moderate-sized datasets

Influence functions: a way to approximate these leave-one-out (LOO) errors

Sample Influence

Q: How does a sample affect the prediction model fit? In particular, what would have happened if we had left out that sample from the model training?

Computationally expensive to leave-one sample out and refit model for even moderate-sized datasets

Influence functions: a way to approximate these leave-one-out (LOO) errors

+ Very classical idea in statistics [Hampel (1972), Cook and Weisberg (1982)]

Sample Influence

Q: How does a sample affect the prediction model fit? In particular, what would have happened if we had left out that sample from the model training?

Computationally expensive to leave-one sample out and refit model for even moderate-sized datasets

Influence functions: a way to approximate these leave-one-out (LOO) errors

- Very classical idea in statistics [Hampel (1972), Cook and Weisberg (1982)]
  - Cook's distance, influence points

Sample Influence

Q: How does a sample affect the prediction model fit? In particular, what would have happened if we had left out that sample from the model training?

Computationally expensive to leave-one sample out and refit model for even moderate-sized datasets

Influence functions: a way to approximate these leave-one-out (LOO) errors

- Very classical idea in statistics [Hampel (1972), Cook and Weisberg (1982)]
  - Cook's distance, influence points
- Recently popularized in modern machine learning [Koh and Liang (2017)]

Influence functions

Consider the following optimization problem

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\mathbf{x}_i, \boldsymbol{\theta}) + \epsilon \mathcal{L}(\mathbf{x}_k, \boldsymbol{\theta})$$

See details in Koh and Liang (2017)

Influence functions

Consider the following optimization problem

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\mathbf{x}_i, \boldsymbol{\theta}) + \epsilon \mathcal{L}(\mathbf{x}_k, \boldsymbol{\theta})$$

When ϵ = 0, this is equivalent to the original fit, trained on all samples

When ϵ = -1/n, this is equivalent to refitting while leaving out sample *k*

See details in Koh and Liang (2017)

Influence functions

Consider the following optimization problem

$$r^{(-k)}(\epsilon) := \arg\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\mathbf{x}_i, \boldsymbol{\theta}) + \epsilon \mathcal{L}(\mathbf{x}_k, \boldsymbol{\theta})$$

When ϵ = 0, this is equivalent to the original fit, trained on all samples

When ϵ = -1/n, this is equivalent to refitting while leaving out sample *k*

See details in Koh and Liang (2017)

Influence functions

Consider the following optimization problem

$$r^{(-k)}(\epsilon) := \arg\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\mathbf{x}_i, \boldsymbol{\theta}) + \epsilon \mathcal{L}(\mathbf{x}_k, \boldsymbol{\theta})$$

When ϵ = 0, this is equivalent to the original fit, trained on all samples

When ϵ = -1/n, this is equivalent to refitting while leaving out sample *k*

Can perform first-order Taylor expansion around ϵ_0 = 0 and plug in ϵ = -1/n to get approximation of LOO fit

See details in Koh and Liang (2017)

Influence functions

Influence of leaving out sample $\boldsymbol{x}_k$ on model parameters:

See details in Koh and Liang (2017)

Influence functions

Influence of leaving out sample $\boldsymbol{x}_k$ on model parameters:

$$\hat{\boldsymbol{\theta}}^{(-k)} - \hat{\boldsymbol{\theta}} = \left(\sum_{i=1}^{n} \nabla^2_{\boldsymbol{\theta}} \mathcal{L}(\mathbf{x}_i, \hat{\boldsymbol{\theta}}) \right)^{-1} \nabla_{\boldsymbol{\theta}} \mathcal{L}(\mathbf{x}_k, \hat{\boldsymbol{\theta}})$$

See details in [Koh and Liang (2017)]

Influence functions

Influence of leaving out sample $\boldsymbol{x}_k$ on model parameters:

$$\hat{\boldsymbol{\theta}}^{(-k)} - \hat{\boldsymbol{\theta}} = \left(\sum_{i=1}^{n} \nabla^2_{\boldsymbol{\theta}} \mathcal{L}(\mathbf{x}_i, \hat{\boldsymbol{\theta}}) \right)^{-1} \nabla_{\boldsymbol{\theta}} \mathcal{L}(\mathbf{x}_k, \hat{\boldsymbol{\theta}})$$

Influence of leaving out sample $\boldsymbol{x}_k$ on the change in loss for sample $\boldsymbol{x}_{test}$:

[application of chain rule]

See details in [Koh and Liang (2017)]

Influence functions

Influence of leaving out sample $\boldsymbol{x}_k$ on model parameters:

$$\hat{\boldsymbol{\theta}}^{(-k)} - \hat{\boldsymbol{\theta}} = \left(\sum_{i=1}^{n} \nabla^2_{\boldsymbol{\theta}} \mathcal{L}(\mathbf{x}_i, \hat{\boldsymbol{\theta}}) \right)^{-1} \nabla_{\boldsymbol{\theta}} \mathcal{L}(\mathbf{x}_k, \hat{\boldsymbol{\theta}})$$

Influence of leaving out sample $\boldsymbol{x}_k$ on the change in loss for sample $\boldsymbol{x}_{test}$:

[application of chain rule]

$$(y_{\text{test}} - \hat{y}^{(-k)}_{\text{test}})^2 - (y_{\text{test}} - \hat{y}_{\text{test}})^2 \approx \nabla_{\boldsymbol{\theta}} \mathcal{L}(\mathbf{x}_{\text{test}}, \hat{\boldsymbol{\theta}})^\top \left(\sum_{i=1}^{n} \nabla^2_{\boldsymbol{\theta}} \mathcal{L}(\mathbf{x}_i, \hat{\boldsymbol{\theta}}) \right)^{-1} \nabla_{\boldsymbol{\theta}} \mathcal{L}(\mathbf{x}_k, \hat{\boldsymbol{\theta}})$$

See details in Koh and Liang (2017)

Influence functions

Influence of leaving out sample $\boldsymbol{x}_k$ on model parameters:

$$\hat{\boldsymbol{\theta}}^{(-k)} - \hat{\boldsymbol{\theta}} = \left(\sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}}^2 \mathcal{L}(\mathbf{x}_i, \hat{\boldsymbol{\theta}}) \right)^{-1} \nabla_{\boldsymbol{\theta}} \mathcal{L}(\mathbf{x}_k, \hat{\boldsymbol{\theta}})$$

Influence of leaving out sample $\boldsymbol{x}_k$ on the change in loss for sample $\boldsymbol{x}_{test}$:

[application of chain rule]

$$(y_{\text{test}} - \hat{y}_{\text{test}}^{(-k)})^2 - (y_{\text{test}} - \hat{y}_{\text{test}})^2 \approx \nabla_{\boldsymbol{\theta}} \mathcal{L}(\mathbf{x}_{\text{test}}, \hat{\boldsymbol{\theta}})^\top \left(\sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}}^2 \mathcal{L}(\mathbf{x}_i, \hat{\boldsymbol{\theta}}) \right)^{-1} \nabla_{\boldsymbol{\theta}} \mathcal{L}(\mathbf{x}_k, \hat{\boldsymbol{\theta}})$$

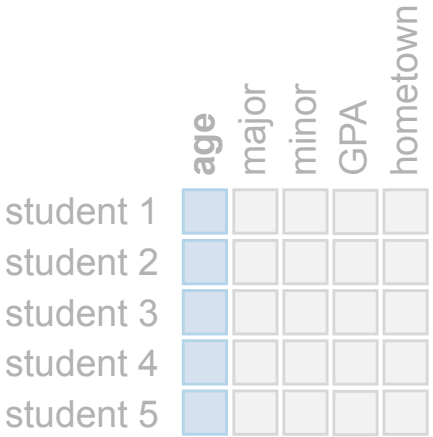
Takeaway: Only need access to gradient and Hessian of the loss function to efficiently approximate LOO effects

See details in Koh and Liang (2017)

(One possible) Taxonomy of interpretability tools

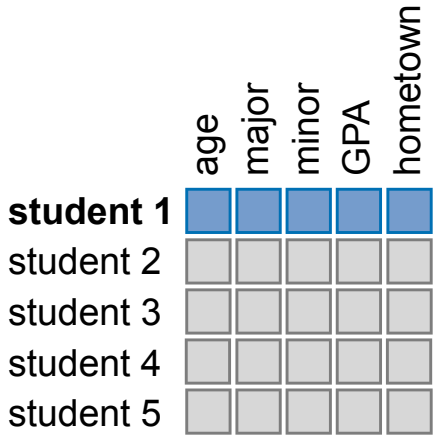
Feature Importances

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



Sample Influences

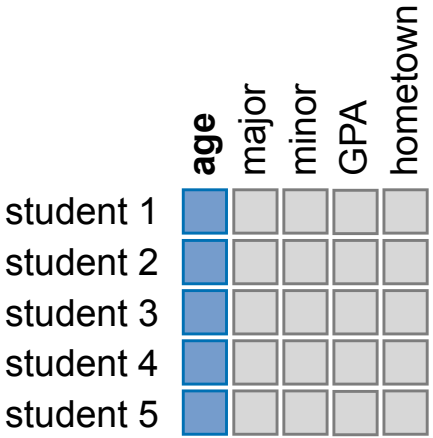
How does leaving out a **sample** change the model fit, predictions, and/or error?



(One possible) Taxonomy of interpretability tools

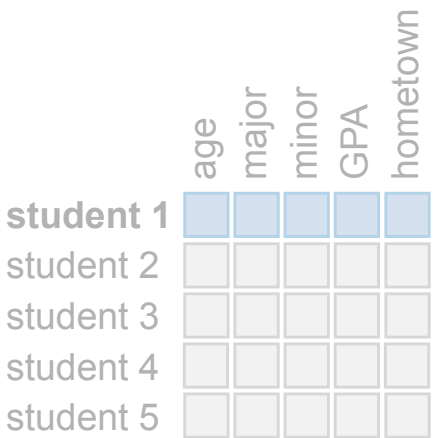
Feature Importances

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?



Feature Importances

(One possible) Taxonomy of interpretability tools

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?

(One possible) Taxonomy of interpretability tools

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?

(One possible) Taxonomy of interpretability tools

Feature Importances

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

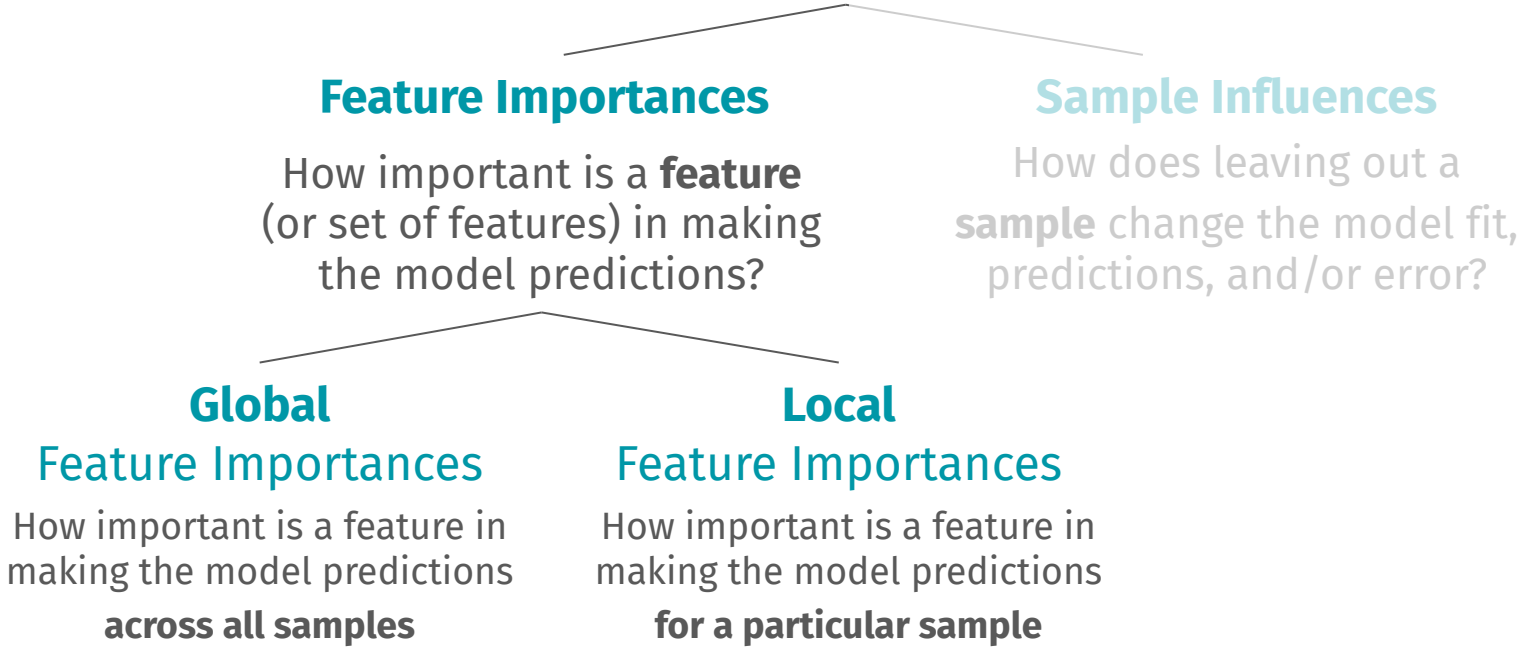
Sample Influences

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

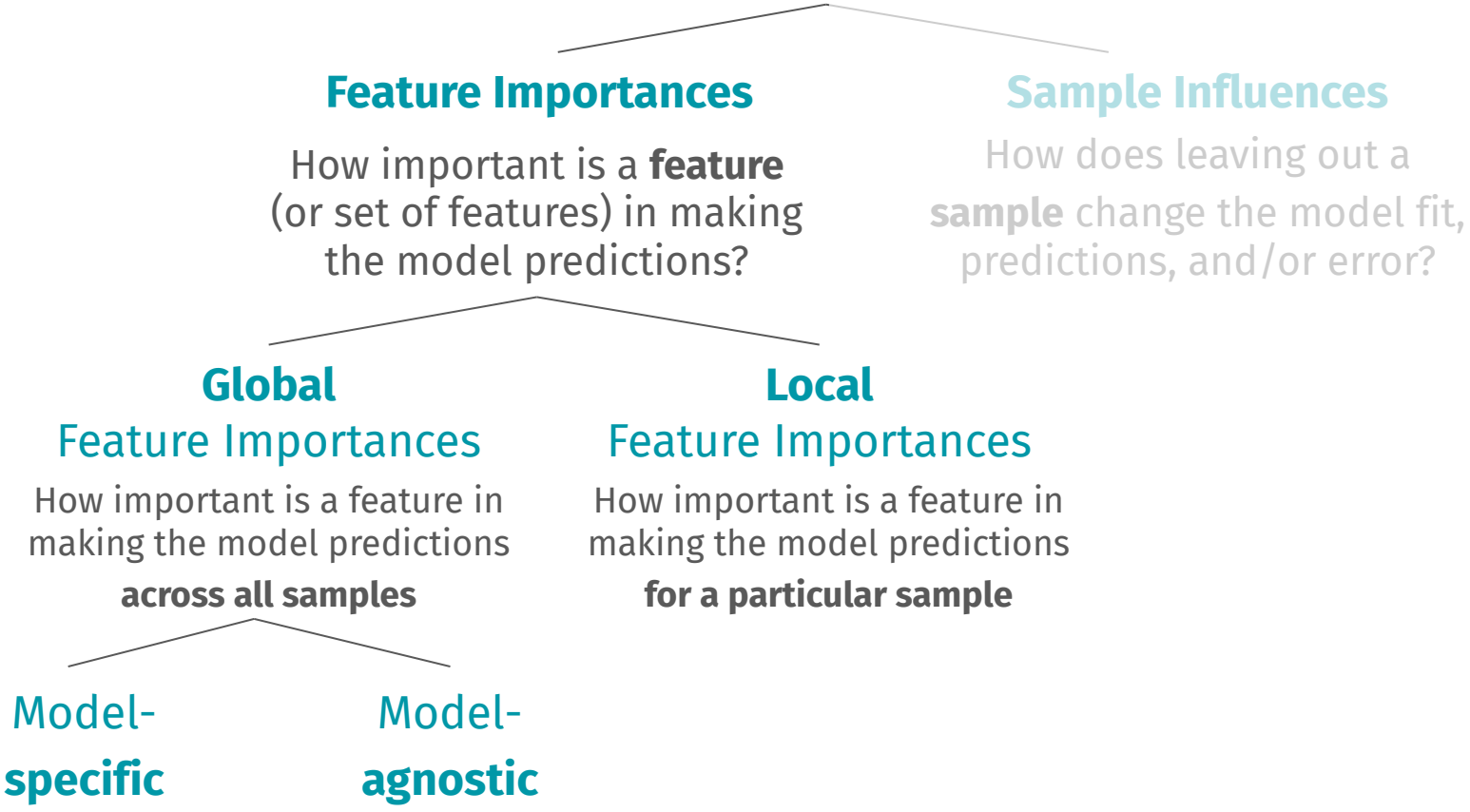
Global
Feature Importances

How important is a feature in making the model predictions **across all samples**

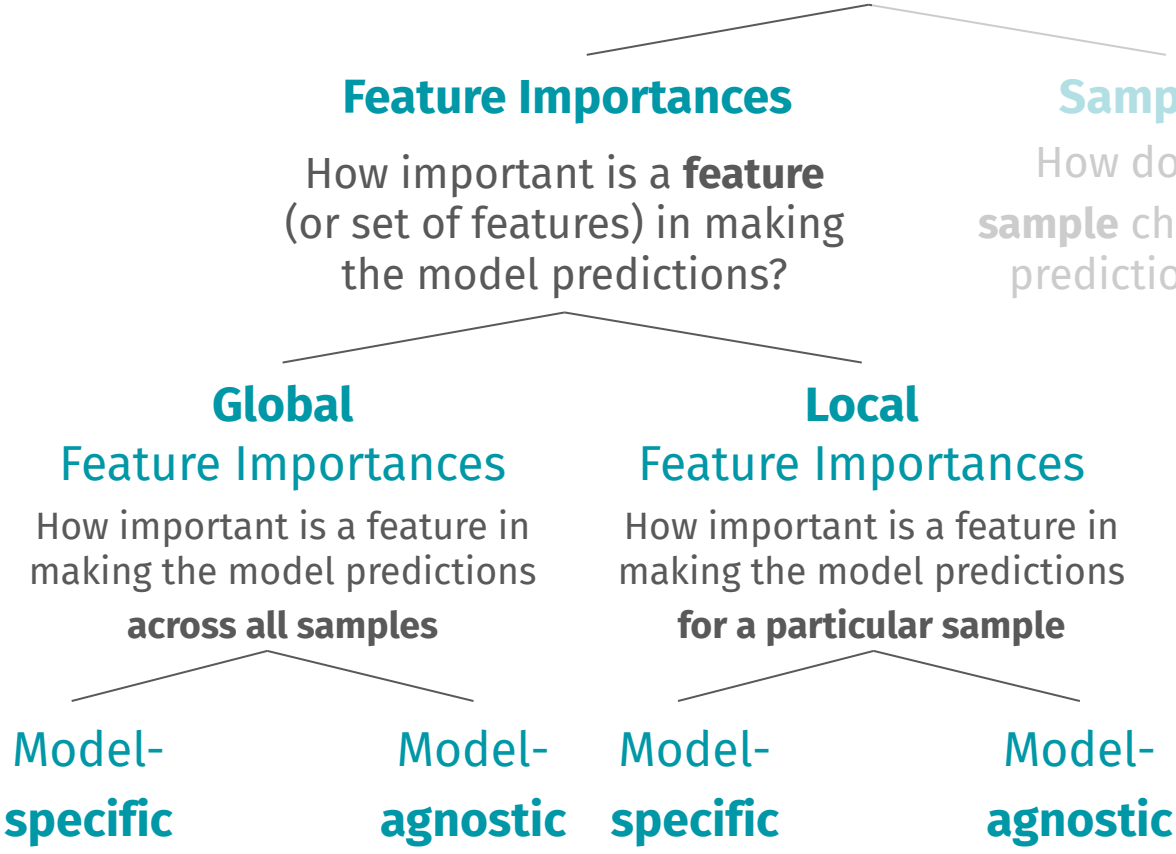
(One possible) Taxonomy of interpretability tools



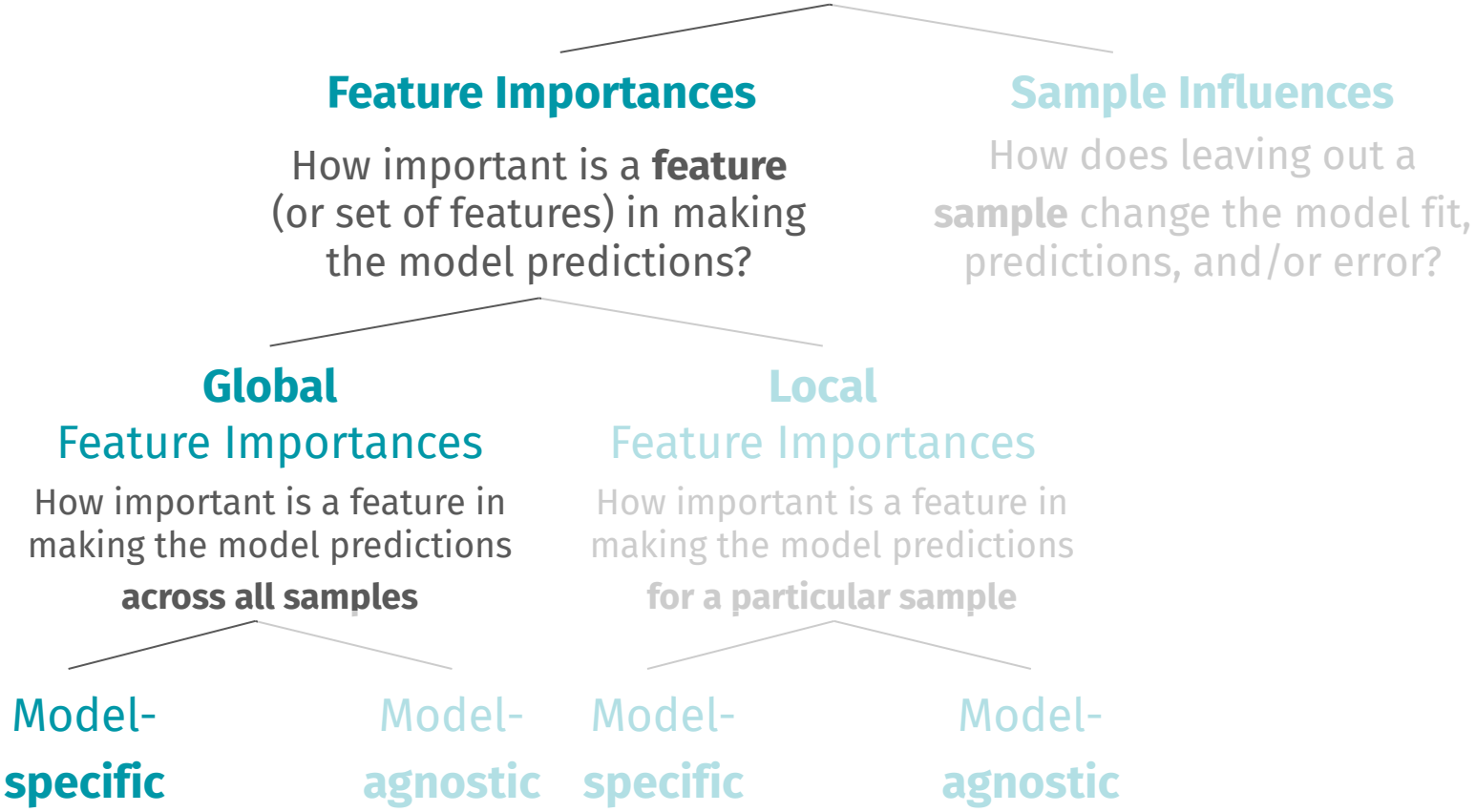
(One possible) Taxonomy of interpretability tools



(One possible) Taxonomy of interpretability tools



(One possible) Taxonomy of interpretations



Overview of Model-specific Global Feature Importances

Overview of Model-specific Global Feature Importances

+ **Linear Regression**

Overview of Model-specific Global Feature Importances

- **Linear Regression**
  - Magnitude of coefficients (if all features in X are measured on same scale)

Overview of Model-specific Global Feature Importances

- **Linear Regression**
  - Magnitude of coefficients (if all features in X are measured on same scale)
  - Z-score/p-value

Overview of Model-specific Global Feature Importances

- **Linear Regression**
  - Magnitude of coefficients (if all features in X are measured on same scale)
  - Z-score/p-value
- **Lasso/ridge/elastic net regression**

Overview of Model-specific Global Feature Importances

- **Linear Regression**
  - Magnitude of coefficients (if all features in X are measured on same scale)
  - Z-score/p-value
- **Lasso/ridge/elastic net regression**
  - Magnitude of coefficients (if all features in X are measured on same scale)

Overview of Model-specific Global Feature Importances

- **Linear Regression**
  - Magnitude of coefficients (if all features in X are measured on same scale)
  - Z-score/p-value
- **Lasso/ridge/elastic net regression**
  - Magnitude of coefficients (if all features in X are measured on same scale)
  - If features are not on same scale, then $|\beta_j|/\text{SD}(X_j)$

Overview of Model-specific Global Feature Importances

- **Linear Regression**
  - Magnitude of coefficients (if all features in X are measured on same scale)
  - Z-score/p-value
- **Lasso/ridge/elastic net regression**
  - Magnitude of coefficients (if all features in X are measured on same scale)
  - If features are not on same scale, then $|\beta_j|/\text{SD}(X_j)$
  - Can we get p-values?

Overview of Model-specific Global Feature Importances

- **Linear Regression**
  - Magnitude of coefficients (if all features in X are measured on same scale)
  - Z-score/p-value
- **Lasso/ridge/elastic net regression**
  - Magnitude of coefficients (if all features in X are measured on same scale)
  - If features are not on same scale, then $|\beta_j|/\text{SD}(X_j)$
  - Can we get p-values?
    - Generally difficult for regularized regression models

Overview of Model-specific Global Feature Importances

- **Linear Regression**
  - Magnitude of coefficients (if all features in X are measured on same scale)
  - Z-score/p-value

- **Lasso/ridge/elastic net regression**
  - Magnitude of coefficients (if all features in X are measured on same scale)
  - If features are not on same scale, then $|\beta_j|/\text{SD}(X_j)$
  - Can we get p-values?
    - Generally difficult for regularized regression models
    - For Lasso, can use (i) *selective inference* or (ii) data splitting (i.e., split 1 is used to fit Lasso and split 2 is used to refit linear regression using only the features with non-zero coefficients from the Lasso fit)

Overview of Model-specific Global Feature Importances

- **Linear Regression**
  - Magnitude of coefficients (if all features in X are measured on same scale)
  - Z-score/p-value
- **Lasso/ridge/elastic net regression**
  - Magnitude of coefficients (if all features in X are measured on same scale)
  - If features are not on same scale, then $|\beta_j|/\text{SD}(X_j)$
  - Can we get p-values?
    - Generally difficult for regularized regression models
    - For Lasso, can use (i) *selective inference* or (ii) data splitting (i.e., split 1 is used to fit Lasso and split 2 is used to refit linear regression using only the features with non-zero coefficients from the Lasso fit)
- **Decision trees/random forest/XGboost**

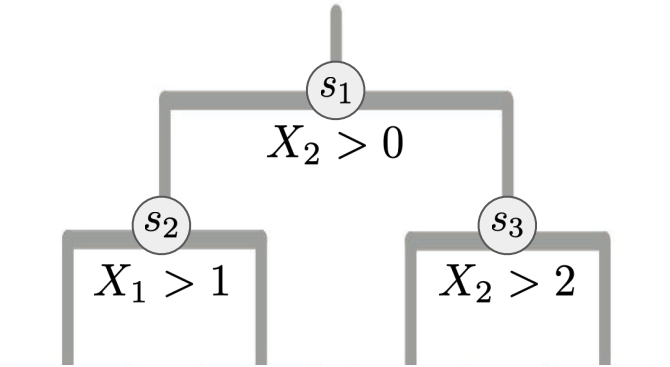
Overview of Model-specific Global Feature Importances

- **Linear Regression**
  - Magnitude of coefficients (if all features in X are measured on same scale)
  - Z-score/p-value
- **Lasso/ridge/elastic net regression**
  - Magnitude of coefficients (if all features in X are measured on same scale)
  - If features are not on same scale, then $|\beta_j|/\text{SD}(X_j)$
  - Can we get p-values?
    - Generally difficult for regularized regression models
    - For Lasso, can use (i) *selective inference* or (ii) data splitting (i.e., split 1 is used to fit Lasso and split 2 is used to refit linear regression using only the features with non-zero coefficients from the Lasso fit)
- **Decision trees/random forest/XGboost**
  - Mean decrease in impurity (MDI)

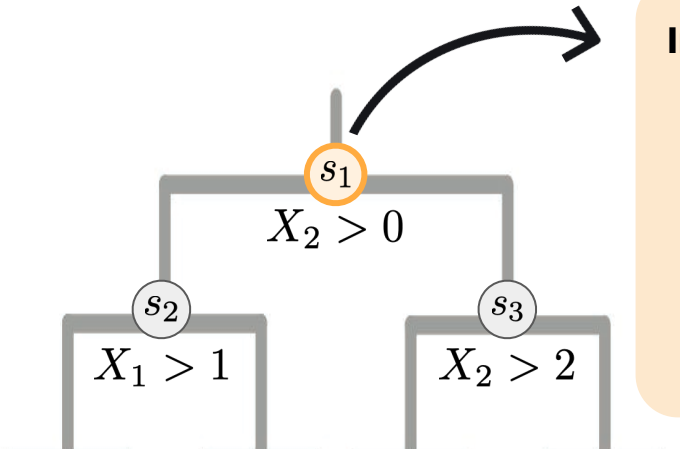
Overview of Model-specific Global Feature Importances

- **Linear Regression**
  - Magnitude of coefficients (if all features in X are measured on same scale)
  - Z-score/p-value
- **Lasso/ridge/elastic net regression**
  - Magnitude of coefficients (if all features in X are measured on same scale)
  - If features are not on same scale, then $|\beta_j|/\text{SD}(X_j)$
  - Can we get p-values?
    - Generally difficult for regularized regression models
    - For Lasso, can use (i) *selective inference* or (ii) data splitting (i.e., split 1 is used to fit Lasso and split 2 is used to refit linear regression using only the features with non-zero coefficients from the Lasso fit)
- **Decision trees/random forest/XGboost**
  - Mean decrease in impurity (MDI)
- **Many more...**

Mean Decrease in Impurity (MDI)

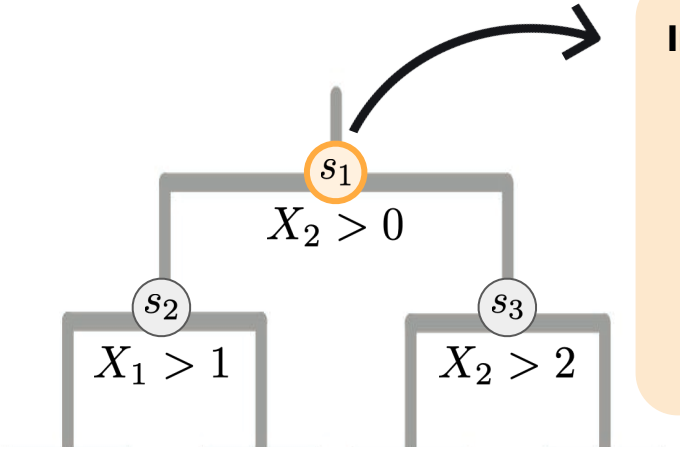


Mean Decrease in Impurity (MDI)



Impurity decrease at s_1

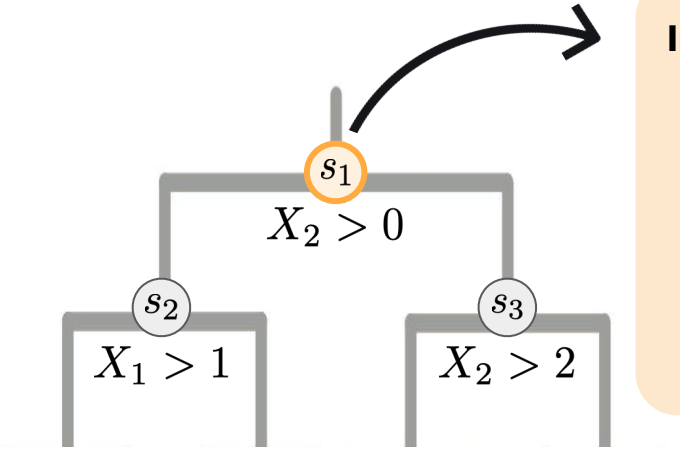
Mean Decrease in Impurity (MDI)



Impurity decrease at s_1

"Measures decrease in variance from making the split"

Mean Decrease in Impurity (MDI)

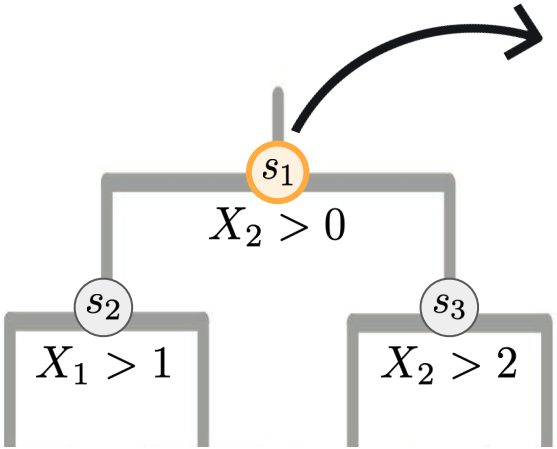


Impurity decrease at s_1

"Measures decrease in variance from making the split"

$$\hat{\Delta}(s_1) = \underbrace{\sum_{\mathbf{x} \in s_1} (y_i - \bar{y}_{s_1})^2}_{\text{Var(node of interest)}} - \underbrace{\sum_{\mathbf{x} \in s_2} (y_i - \bar{y}_{s_2})^2}_{\text{Var(left child node)}} - \underbrace{\sum_{\mathbf{x} \in s_3} (y_i - \bar{y}_{s_3})^2}_{\text{Var(right child node)}}$$

Mean Decrease in Impurity (MDI)



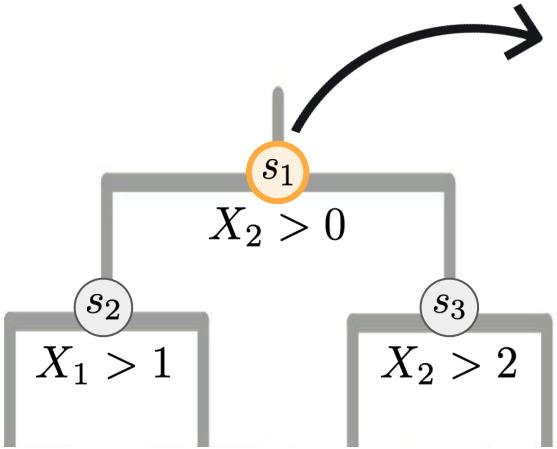
Impurity decrease at s_1

"Measures decrease in variance from making the split"

$$\hat{\Delta}(s_1) = \underbrace{\sum_{\mathbf{x} \in s_1} (y_i - \bar{y}_{s_1})^2}_{\text{Var(node of interest)}} - \underbrace{\sum_{\mathbf{x} \in s_2} (y_i - \bar{y}_{s_2})^2}_{\text{Var(left child node)}} - \underbrace{\sum_{\mathbf{x} \in s_3} (y_i - \bar{y}_{s_3})^2}_{\text{Var(right child node)}}$$

For each feature k, MDI(k) is the weighted sum of impurity decreases across nodes that split on X_k, e.g.,

Mean Decrease in Impurity (MDI)



Impurity decrease at s_1

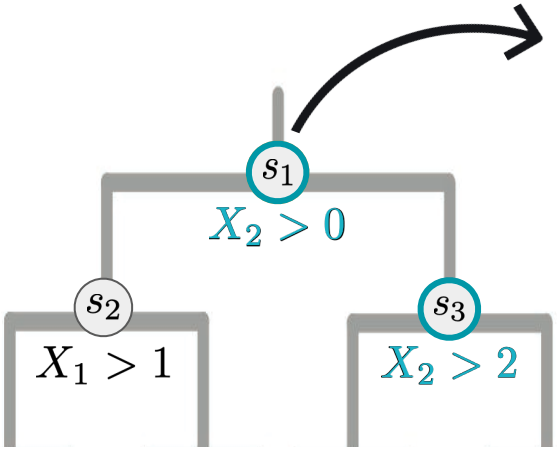
"Measures decrease in variance from making the split"

$$\hat{\Delta}(s_1) = \underbrace{\sum_{\mathbf{x} \in s_1} (y_i - \bar{y}_{s_1})^2}_{\text{Var(node of interest)}} - \underbrace{\sum_{\mathbf{x} \in s_2} (y_i - \bar{y}_{s_2})^2}_{\text{Var(left child node)}} - \underbrace{\sum_{\mathbf{x} \in s_3} (y_i - \bar{y}_{s_3})^2}_{\text{Var(right child node)}}$$

For each feature k, MDI(k) is the weighted sum of impurity decreases across nodes that split on X_k, e.g.,

$$MDI(X_2) = \frac{n_1}{n}\hat{\Delta}(s_1) + \frac{n_3}{n}\hat{\Delta}(s_3)$$

Mean Decrease in Impurity (MDI)



Impurity decrease at s_1

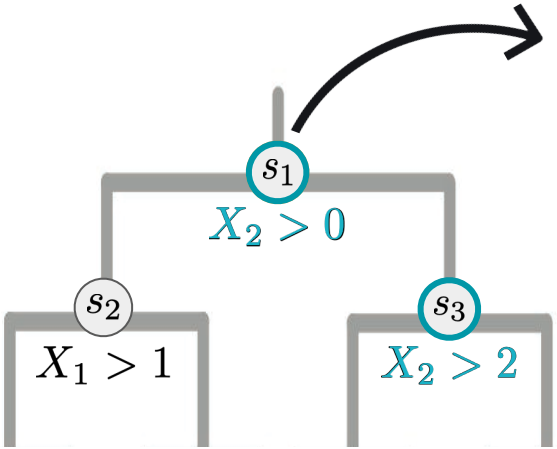
"Measures decrease in variance from making the split"

$$\hat{\Delta}(s_1) = \underbrace{\sum_{\mathbf{x} \in s_1} (y_i - \bar{y}_{s_1})^2}_{\text{Var(node of interest)}} - \underbrace{\sum_{\mathbf{x} \in s_2} (y_i - \bar{y}_{s_2})^2}_{\text{Var(left child node)}} - \underbrace{\sum_{\mathbf{x} \in s_3} (y_i - \bar{y}_{s_3})^2}_{\text{Var(right child node)}}$$

For each feature k, MDI(k) is the weighted sum of impurity decreases across nodes that split on X_k, e.g.,

$$MDI(X_2) = \frac{n_1}{n}\hat{\Delta}(s_1) + \frac{n_3}{n}\hat{\Delta}(s_3)$$

Mean Decrease in Impurity (MDI)



Impurity decrease at s_1

"Measures decrease in variance from making the split"

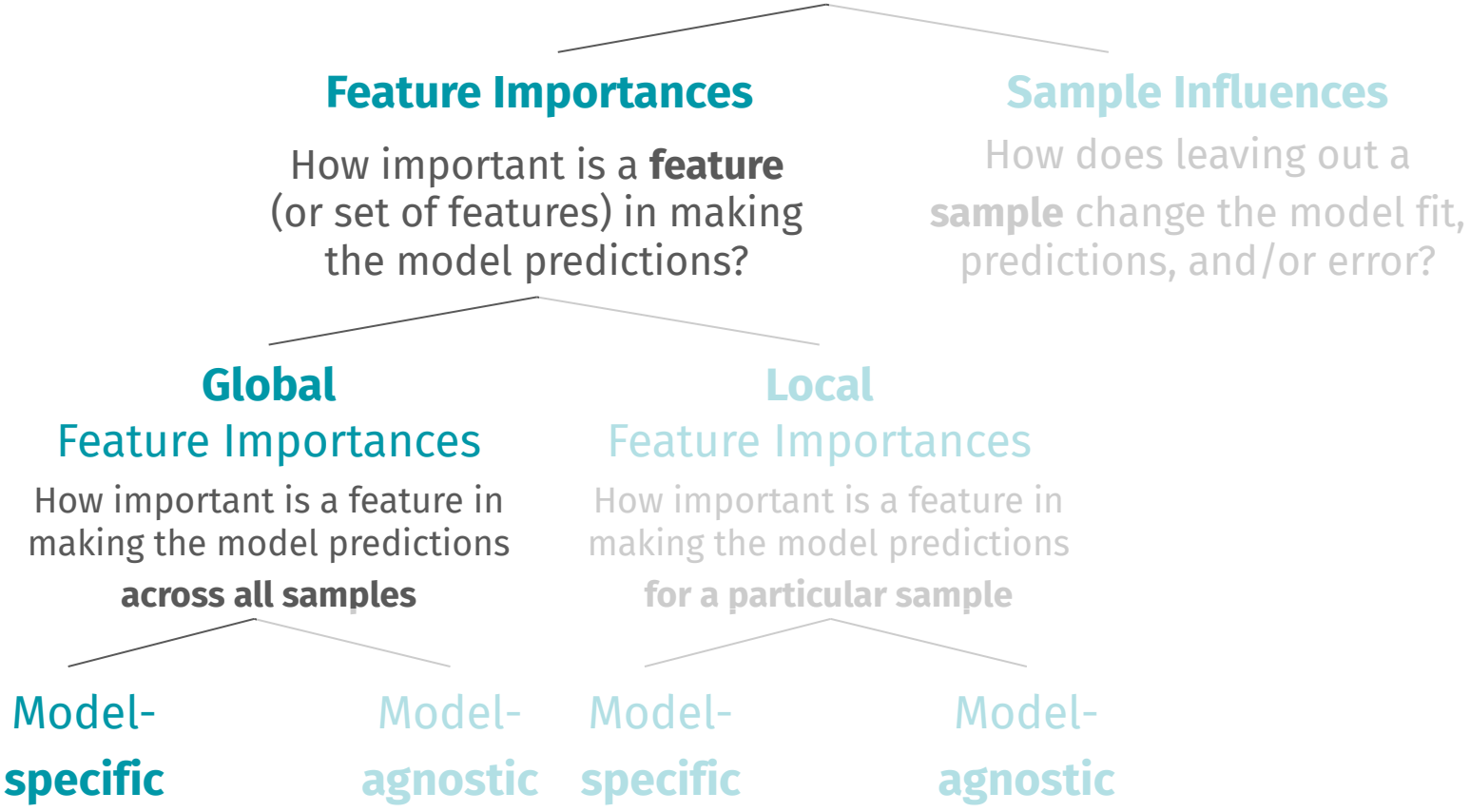
$$\hat{\Delta}(s_1) = \underbrace{\sum_{\mathbf{x} \in s_1} (y_i - \bar{y}_{s_1})^2}_{\text{Var(node of interest)}} - \underbrace{\sum_{\mathbf{x} \in s_2} (y_i - \bar{y}_{s_2})^2}_{\text{Var(left child node)}} - \underbrace{\sum_{\mathbf{x} \in s_3} (y_i - \bar{y}_{s_3})^2}_{\text{Var(right child node)}}$$

For each feature k, MDI(k) is the weighted sum of impurity decreases across nodes that split on X_k, e.g.,

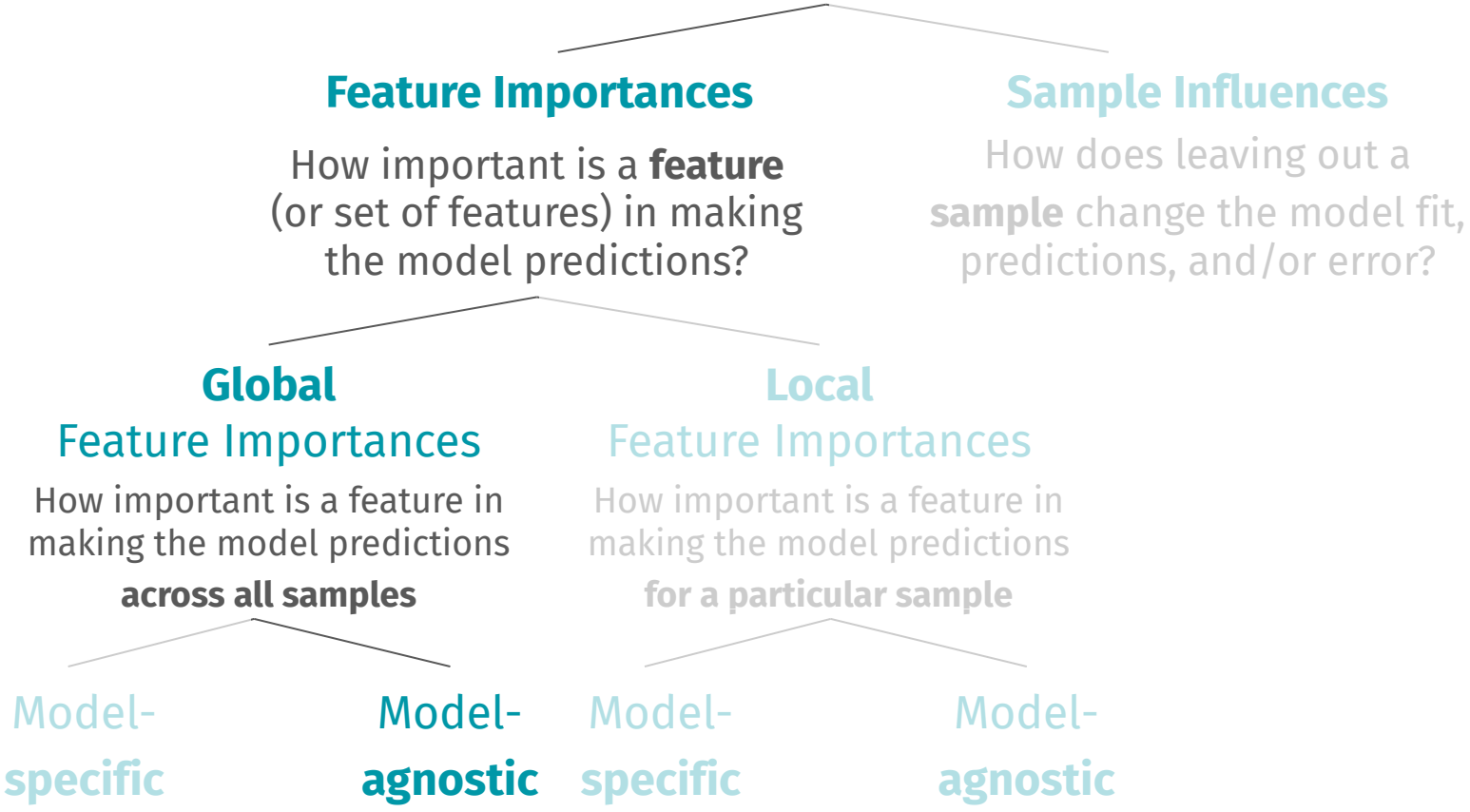
$$MDI(X_2) = \frac{n_1}{n}\hat{\Delta}(s_1) + \frac{n_3}{n}\hat{\Delta}(s_3)$$

MDI(k) for a forest is the average MDI(k) across all trees

(One possible) Taxonomy of interpretations



(One possible) Taxonomy of interpretations



Overview of Model-agnostic Global Feature Importances

- **Feature permutation importance**
- **Feature occlusion or leave-one-covariate out (LOCO) importance**
- **SHAP (SHapley Additive exPlanations)**

Overview of Model-agnostic Global Feature Importances

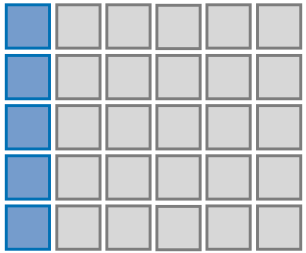
- **Feature permutation importance**
- **Feature occlusion or leave-one-covariate out (LOCO) importance**
- **SHAP (SHapley Additive exPlanations)**

Feature **permutation** importance

- Measures change in model's prediction error between observed prediction error and prediction error from input with feature *k* permuted
- If prediction error from permuted data is much lower than original prediction error, then that feature (the one being permuted) is highly important

Given model *f*,

Original data matrix



Permuted data matrix

$$Perm\ Imp_1 = \frac{1}{B}\sum_{b=1}^{B}\left(\text{prediction error} - \text{prediction error}\right)$$

[i.e., average across B permutations]

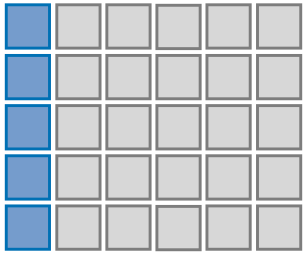
Feature **permutation** importance

- Measures change in model's prediction error between observed prediction error and prediction error from input with feature *k* permuted
- If prediction error from permuted data is much lower than original prediction error, then that feature (the one being permuted) is highly important

Given model *f*,

(Typically better to evaluate predictions on non-training data)

Original data matrix



model *f*

Permuted data matrix

model *f*

$$\textit{Perm Imp}_1 = \frac{1}{B}\sum_{b=1}^{B}\left(\text{prediction error} - \text{prediction error}\right)$$

[i.e., average across B permutations]

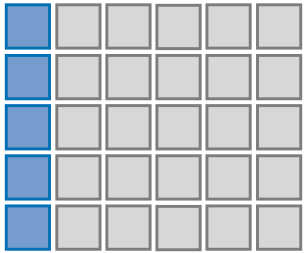
Feature **permutation** importance

- Measures change in model's prediction error between observed prediction error and prediction error from input with feature *k* permuted
- If prediction error from permuted data is much lower than original prediction error, then that feature (the one being permuted) is highly important

Given model *f*,

(Typically better to evaluate predictions on non-training data)

Original data matrix



model *f*

Permuted data matrix

model *f*

$$Perm\ Imp_1 = \frac{1}{B}\sum_{b=1}^{B}\left(\text{prediction error} - \text{prediction error}\right)$$

[i.e., average across B permutations]

Repeat for each feature *k* separately

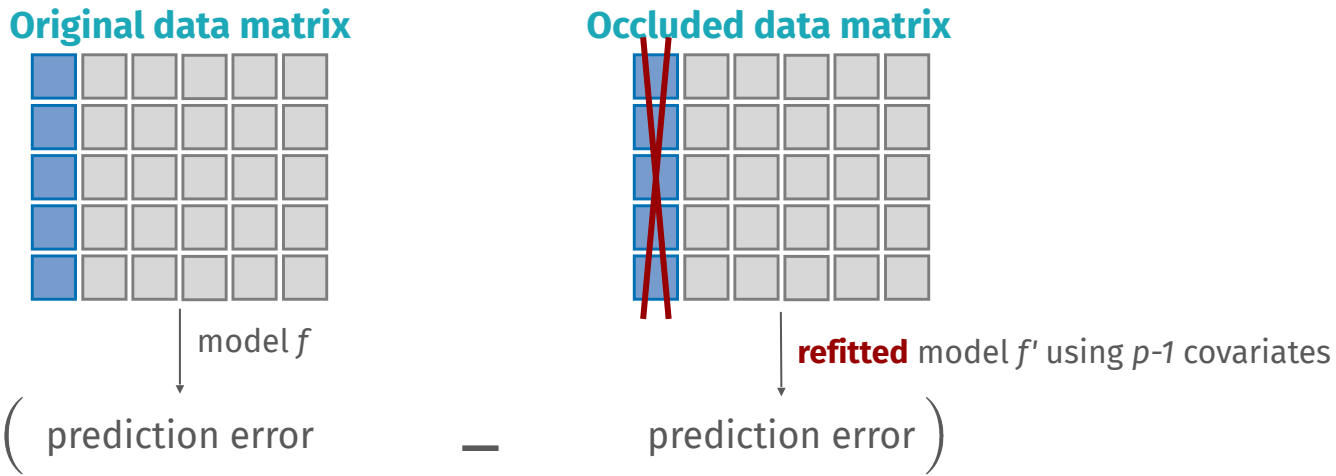
Model-agnostic global feature importances

- **Feature permutation importance**
  - Measures change in prediction error between observed prediction error and prediction error from input with feature *k* permuted
- **Feature occlusion or leave-one-covariate out (LOCO) importance**
- **SHAP (SHapley Additive exPlanations)**

Feature **occlusion** importance (or **leave-one-covariate-out** (LOCO))

- Measures change in model's prediction error between observed prediction error and prediction error from model that has been refitted on data with feature k omitted
- If prediction error from occluded data is much lower than original prediction error, then that feature (the one being occluded) is highly important

Given model *f*,



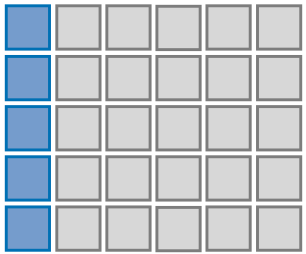
Feature **occlusion** importance (or **leave-one-covariate-out** (LOCO))

- Measures change in model's prediction error between observed prediction error and prediction error from model that has been refitted on data with feature k omitted
- If prediction error from occluded data is much lower than original prediction error, then that feature (the one being occluded) is highly important

Given model *f*,

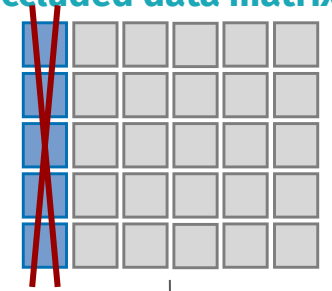
(Typically better to evaluate predictions on non-training data)

Original data matrix



model *f*

Occluded data matrix



refitted model *f'* using *p-1* covariates

$$LOCO\ Imp_1 = \left(\text{prediction error}\right) - \left(\text{prediction error}\right)$$

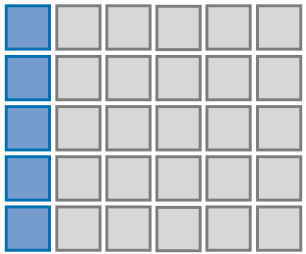
Feature **occlusion** importance (or **leave-one-covariate-out** (LOCO))

- Measures change in model's prediction error between observed prediction error and prediction error from model that has been refitted on data with feature k omitted
- If prediction error from occluded data is much lower than original prediction error, then that feature (the one being occluded) is highly important

Given model *f*,

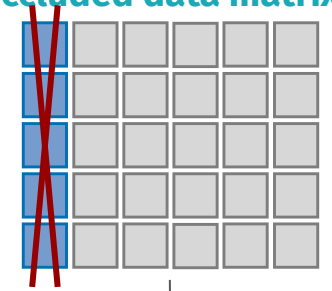
(Typically better to evaluate predictions on non-training data)

Original data matrix



model *f*

Occluded data matrix



refitted model *f'* using *p-1* covariates

$$LOCO\ Imp_1 = \left(\text{prediction error} \right) - \left(\text{prediction error} \right)$$

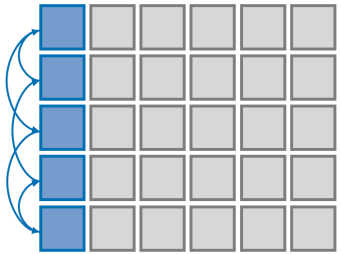
Repeat for each feature *k* separately

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

To measure importance of feature *k*,

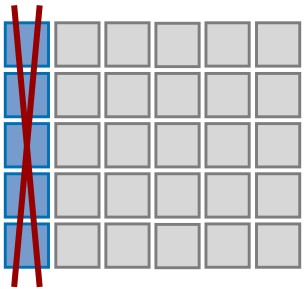
Feature Permutation

- 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*



Model-agnostic global feature importances

Suppose we want to evaluate the global feature importance of feature *k*

- **Feature permutation importance**
  - Measures change in prediction error between observed prediction error and prediction error from input with feature *k* permuted
- **Feature occlusion or leave-one-covariate out (LOCO) importance**
  - Measures change in prediction error between observed prediction error and prediction error from model that has been refitted on data without feature *k*
- **SHAP (SHapley Additive exPlanations)**
  - Approximation of the average marginal contribution of feature *k* to the predictions across all possible combinations of features
  - Details next time...

Partial Dependence Plots (PDP)

Partial Dependence Plots (PDP)

- Many methods (e.g., RF MDI, permutation importance, LOCO importance, …) do not automatically provide the **direction** of the feature's importance
  - Q: Do larger feature values lead to a smaller or lower predicted value?
  - Q: Is the relationship linear or nonlinear?

Partial Dependence Plots (PDP)

- Many methods (e.g., RF MDI, permutation importance, LOCO importance, ...) do not automatically provide the **direction** of the feature's importance
  - Q: Do larger feature values lead to a smaller or lower predicted value?
  - Q: Is the relationship linear or nonlinear?
- **Partial dependence plots** approximate the **marginal effect** of a feature across a range of values on the predicted outcome from a model

Partial Dependence Plots (PDP)

- Many methods (e.g., RF MDI, permutation importance, LOCO importance, ...) do not automatically provide the **direction** of the feature's importance
  - Q: Do larger feature values lead to a smaller or lower predicted value?
  - Q: Is the relationship linear or nonlinear?
- **Partial dependence plots** approximate the **marginal effect** of a feature across a range of values on the predicted outcome from a model

$$\hat{f}_S(x_S) = E_{X_C}\left[\hat{f}(x_S, X_C)\right] = \int \hat{f}(x_S, X_C) d\mathbb{P}(X_C)$$

Partial Dependence Plots (PDP)

- Many methods (e.g., RF MDI, permutation importance, LOCO importance, ...) do not automatically provide the **direction** of the feature's importance
  - Q: Do larger feature values lead to a smaller or lower predicted value?
  - Q: Is the relationship linear or nonlinear?
- **Partial dependence plots** approximate the **marginal effect** of a feature across a range of values on the predicted outcome from a model

$$\hat{f}_S(x_S) = E_{X_C}\left[\hat{f}(x_S, X_C)\right] = \int \hat{f}(x_S, X_C) d\mathbb{P}(X_C)$$

Feature of interest

Partial Dependence Plots (PDP)

- Many methods (e.g., RF MDI, permutation importance, LOCO importance, ...) do not automatically provide the **direction** of the feature's importance
  - Q: Do larger feature values lead to a smaller or lower predicted value?
  - Q: Is the relationship linear or nonlinear?
- **Partial dependence plots** approximate the **marginal effect** of a feature across a range of values on the predicted outcome from a model

$$\hat{f}_S(x_S) = E_{X_C}\left[\hat{f}(x_S, X_C)\right] = \int \hat{f}(x_S, X_C) d\mathbb{P}(X_C)$$

Feature of interest

All other features

Partial Dependence Plots (PDP)

- Many methods (e.g., RF MDI, permutation importance, LOCO importance, ...) do not automatically provide the **direction** of the feature's importance
  - Q: Do larger feature values lead to a smaller or lower predicted value?
  - Q: Is the relationship linear or nonlinear?
- **Partial dependence plots** approximate the **marginal effect** of a feature across a range of values on the predicted outcome from a model

$$\hat{f}_S(x_S) = E_{X_C}\left[\hat{f}(x_S, X_C)\right] = \int \hat{f}(x_S, X_C) d\mathbb{P}(X_C)$$

Feature of interest — All other features

How to estimate integral in practice?

Partial Dependence Plots (PDP)

- Many methods (e.g., RF MDI, permutation importance, LOCO importance, ...) do not automatically provide the **direction** of the feature's importance
  - Q: Do larger feature values lead to a smaller or lower predicted value?
  - Q: Is the relationship linear or nonlinear?
- **Partial dependence plots** approximate the **marginal effect** of a feature across a range of values on the predicted outcome from a model

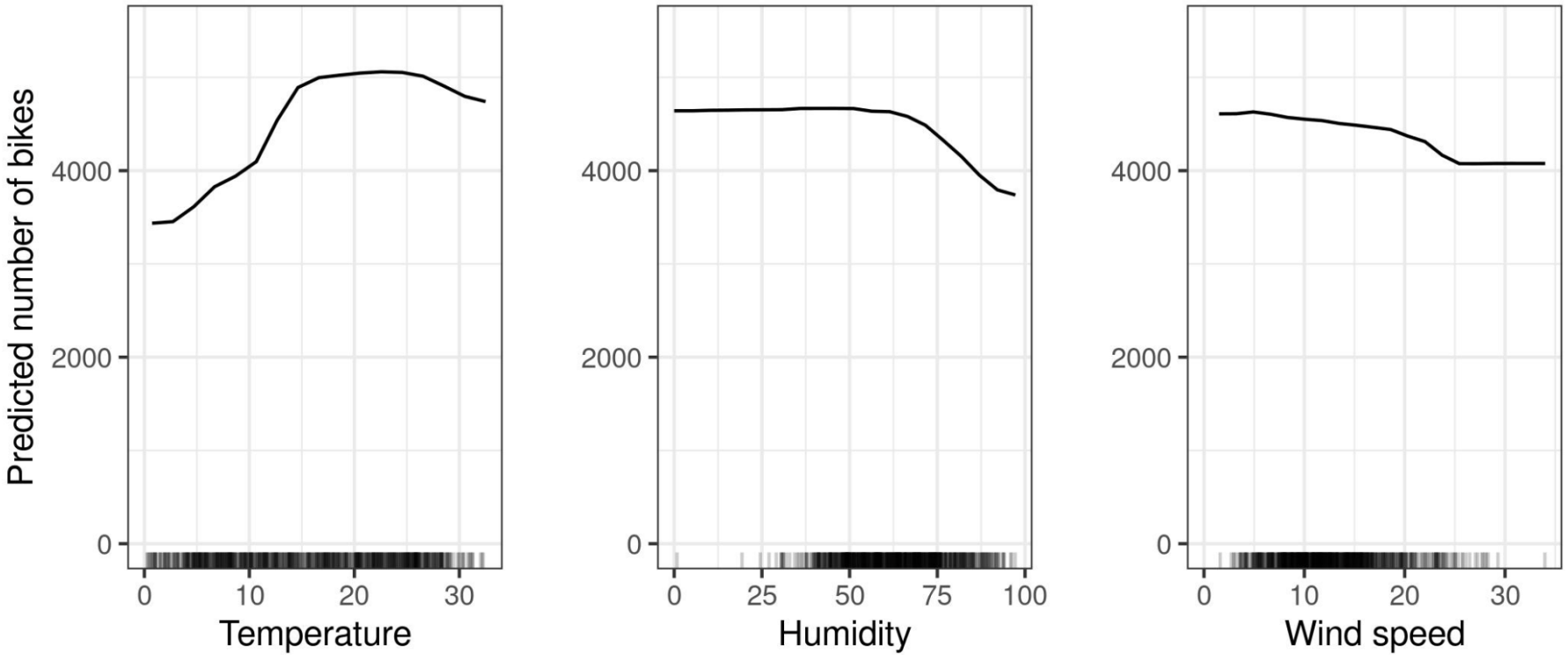
$$\hat{f}_S(x_S) = E_{X_C}\left[\hat{f}(x_S, X_C)\right] = \int \hat{f}(x_S, X_C) d\mathbb{P}(X_C)$$

Feature of interest — All other features

How to estimate integral in practice? $\hat{f}_S(x_S) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(x_S, x_C^{(i)})$

Partial Dependence Plots (PDP)

$$\hat{f}_S(x_S) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}\left(x_S, x_C^{(i)}\right)$$



Next Time

For more on interpretable machine learning, check out this interpretable machine learning textbook by Christoph Molnar