Explainable Models
A Primer

By

Hardeep Arora

Based on book - https://christophm.github.io/interpretable-ml-book/

About Me

18+ years experience in Banking and Finance (MBA Finance, B.E. Computers)

Worked as Developer, DBA, Data Scientist, and Leading Analytics functions at Bank.

Currently Heading COE AI/Analytics @ Accenture.

Advisor for some startups on AI Strategy

Work Life : Credit Risk and Financial Crime

Night Life : Deep Learning, Blockchain, AGI, Kaggle

When is Explainability important?

- Cases where **WHY** is more important than **WHAT,** i.e. when the model has a significant impact.
  - Examples
    - Why was my loan rejected?
    - Did the treatment work?
- Problem domain is not well studied.
- Models don’t work the way we understand the world as humans.
- Understanding **why** help forward the scientific endeavors.

What are the key thing you want to explain?

- **Algorithm**
- **Model as a whole**
- **Parts of a model**
- **Few specific prediction in a model**
- **Group of predictions in a model**

Dataset 1 – Bike Sharing (Regression)

- season : spring (1), summer (2), autumn (3), winter (4).
- holiday : Binary feature indicating if the day was a holiday (1) or not (0).
- yr: The year (2011 or 2012).
- days_since_2011: Number of days since the 01.01.2011 (the first day in the dataset). This feature was introduced to account for the trend, in this case that the bike rental service became more popular over time.
- workingday : Binary feature indicating if the day was a workingday (1) or weekend / holiday (0).
- weathersit : The weather situation on that day
  - Clear, Few clouds, Partly cloudy, Cloudy
  - Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
  - Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
  - Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- temp : Temperature in degrees Celsius.
- hum: Relative humidity in percent (0 to 100).
- windspeed: Wind speed in km per hour.
- cnt: Count of total rental bikes including both casual and registered. The count was used as the target in the regression tasks.

Dataset 2 – Risk Factors for Cervical Cancer (Classification)

- Age in years
- Number of sexual partners
- First sexual intercourse (age in years)
- Number of pregnancies
- Smokes yes (1) or no (1)
- Smokes (years)
- Hormonal Contraceptives yes (1) or no (0)
- Hormonal Contraceptives (years)
- IUD: Intrauterine device yes (1) or no (1)
- IUD (years): Number of years with an intrauterine device
- STDs: Ever had a sexually transmitted disease? Yes (1) or no (0)
- STDs (number): Number of sexually transmitted diseases.
- STDs: Number of diagnosis
- STDs: Time since first diagnosis
- STDs: Time since last diagnosis
- Biopsy: Biopsy results “Healthy” or “Cancer”. Target outcome.

Key Approaches of model explainability

Intrinsically Explainable Models

Model itself is simple and explainable, example

- Short decision tree or
- Linear regression

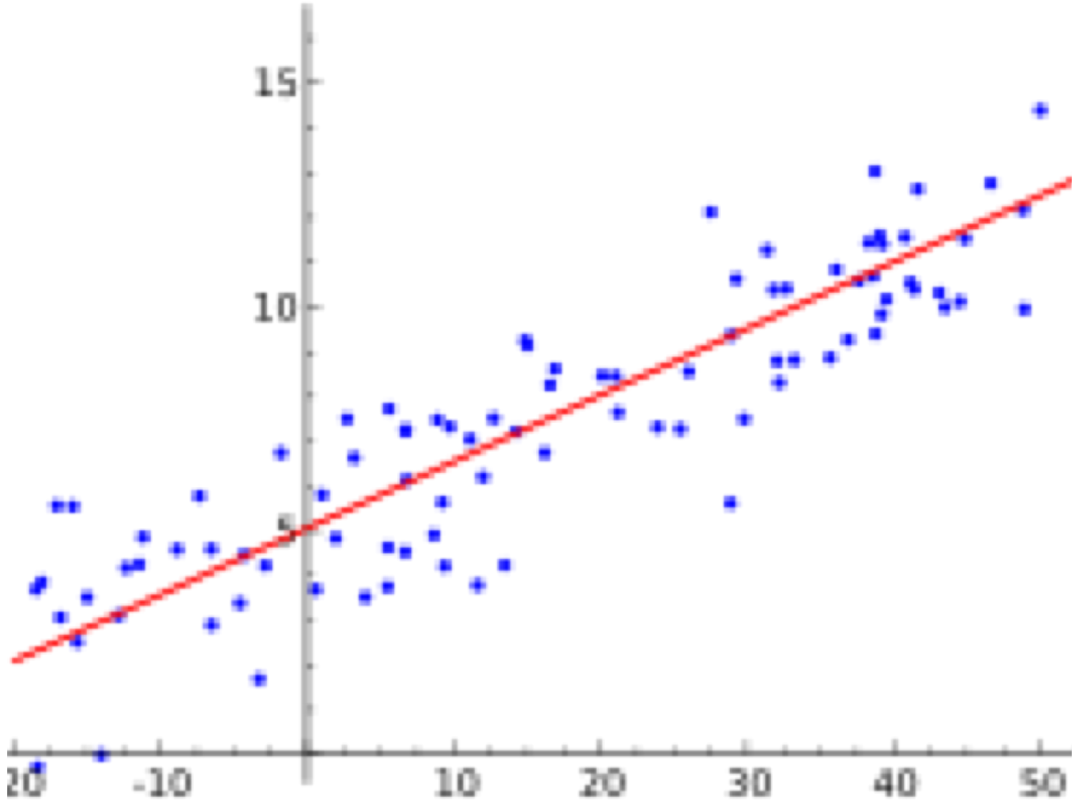
Model Agnostic Methods

Treat model as black box and apply methods from outside to explain them.

Intrinsically Explainable Models

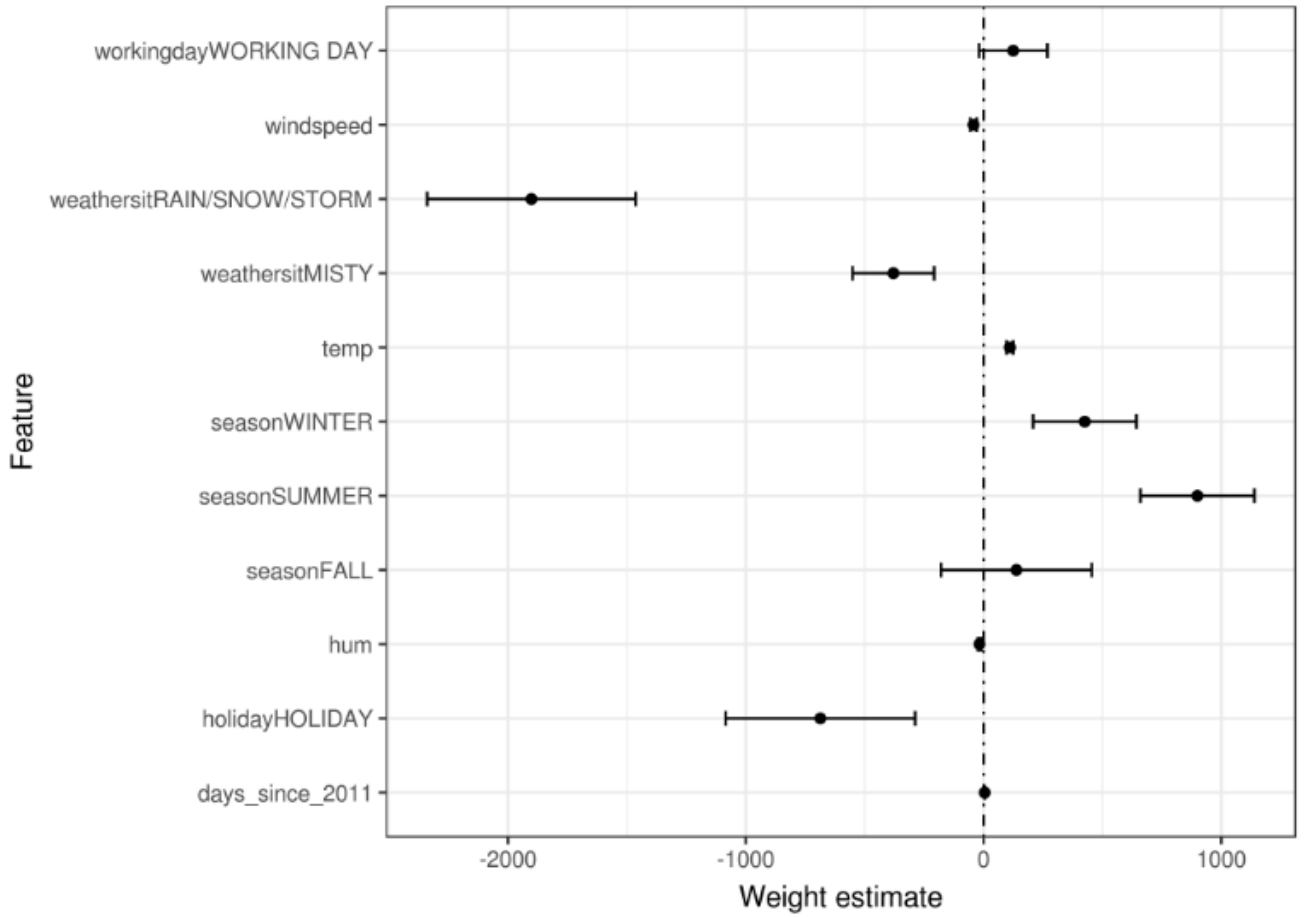
Algorithm	Linear	Monotone	Interaction	Task
Linear models	Yes	Yes	No	Regr.
Logistic regression	No	Yes	No	Class.
Decision trees	No	No	Yes	Class. + Regr.
RuleFit	Yes	No	Yes	Class. + Regr.
Naive Bayes	Yes	Yes	No	Class.n
k-nearest neighbours	No	No	No	Class. + Regr.

Linear Model -Interpretability

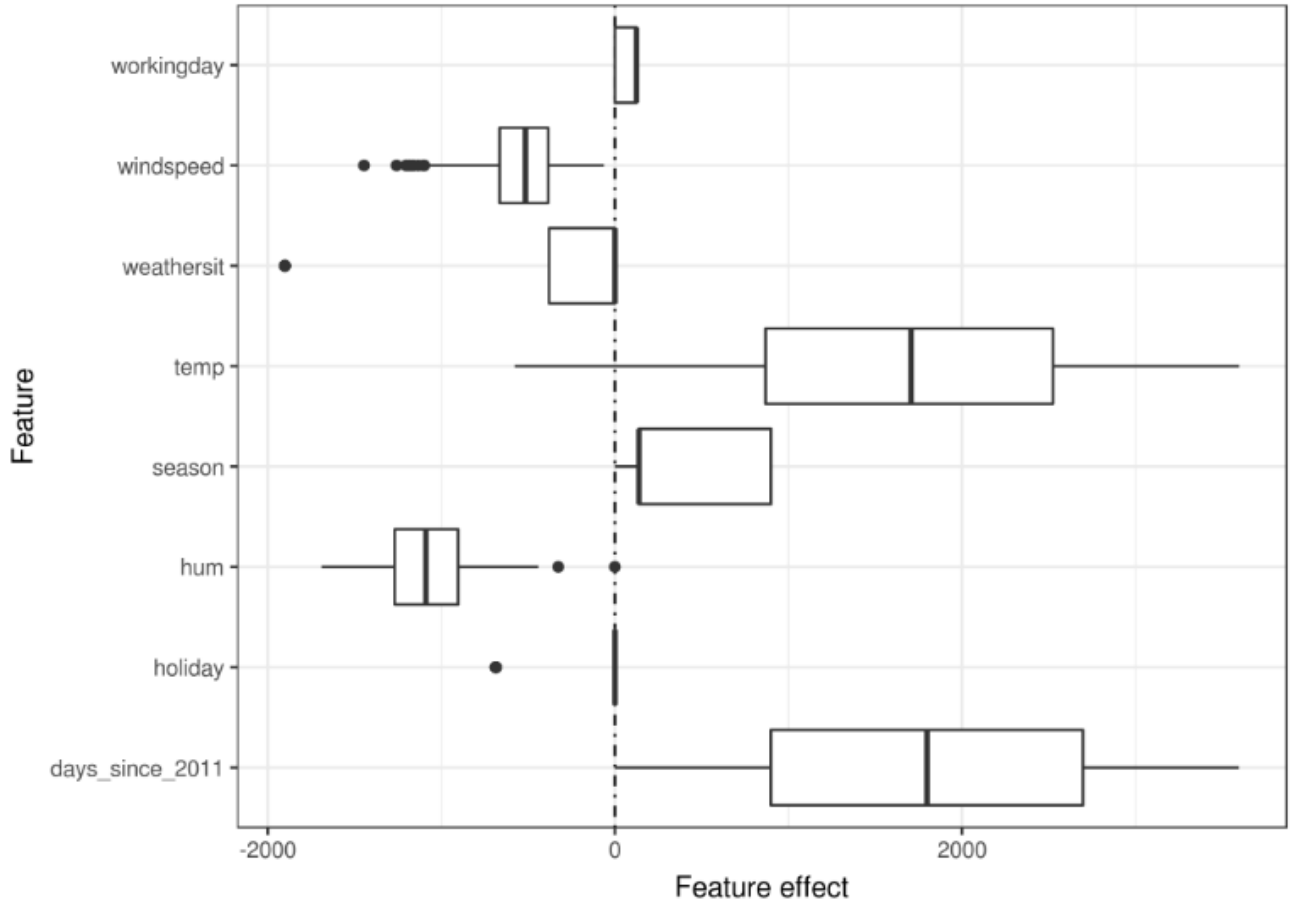


	Weight estimate	Std. Error
(Intercept)	2399.4	238.3
seasonSUMMER	899.3	122.3
seasonFALL	138.2	161.7
seasonWINTER	425.6	110.8
holidayHOLIDAY	-686.1	203.3
workingdayWORKING DAY	124.9	73.3
weathersitMISTY	-379.4	87.6
weathersitRAIN/SNOW/STORM	-1901.5	223.6
temp	110.7	7.0
hum	-17.4	3.2

Weights Plot

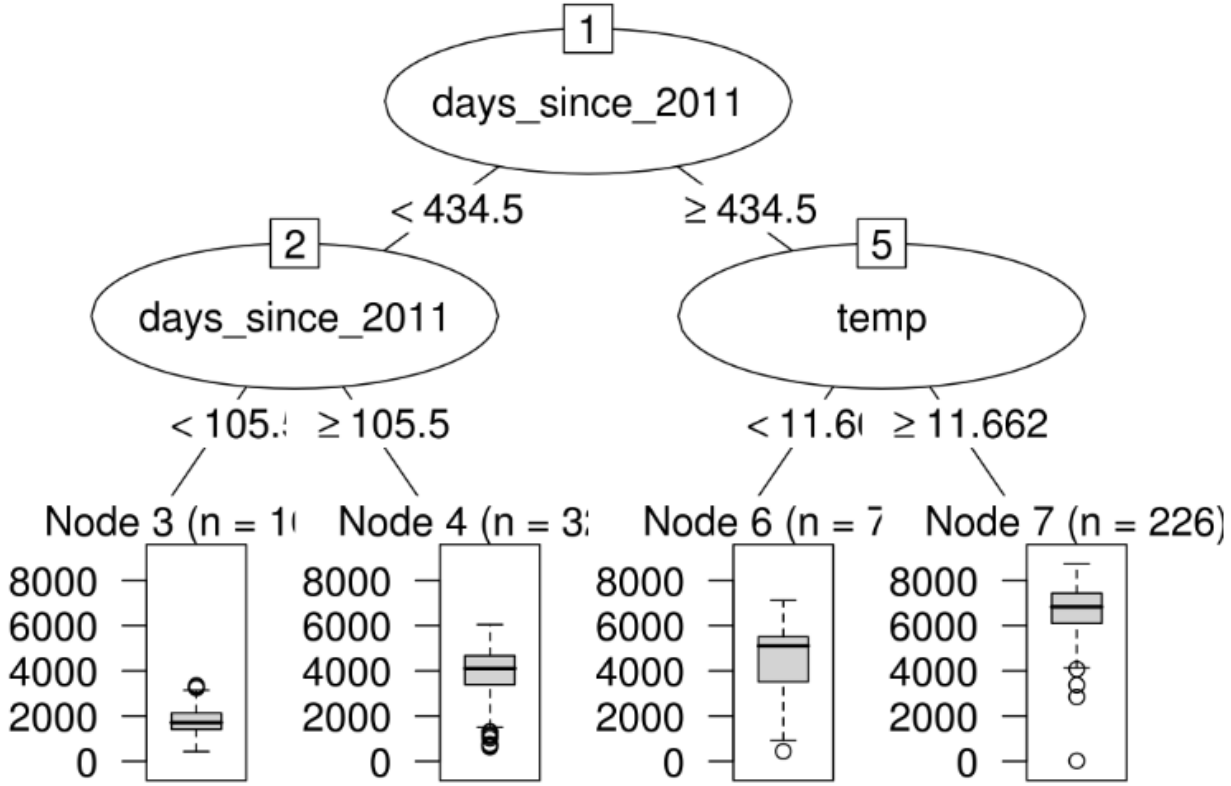


Effects Plot



Decision Trees

- Perfectly suited to **cover interactions** between features in the data
- Has a **natural visualization**, with its nodes and edges
- Trees **create good explanations**



Model-Agnostic Methods

Model flexibility

Works on any underlying model

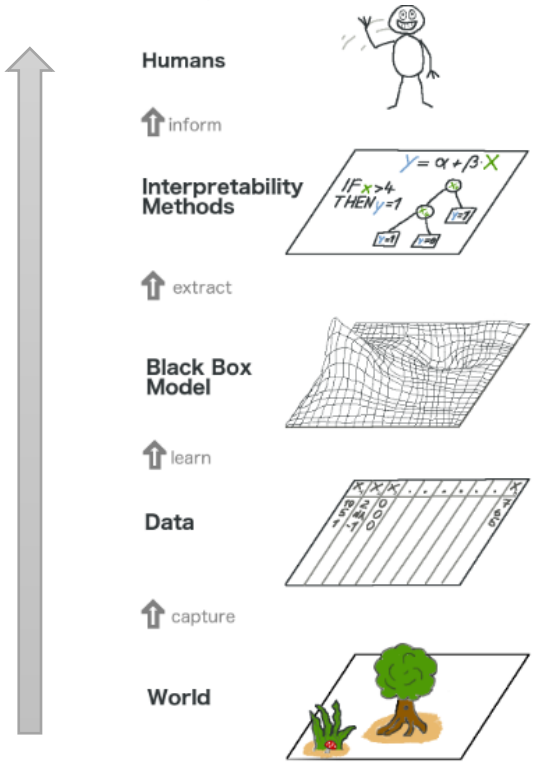
Explanation flexibility

We can use any form of explainability

Representation flexibility

The explain model can use different feature representation from the base model

Where these methods fit ?



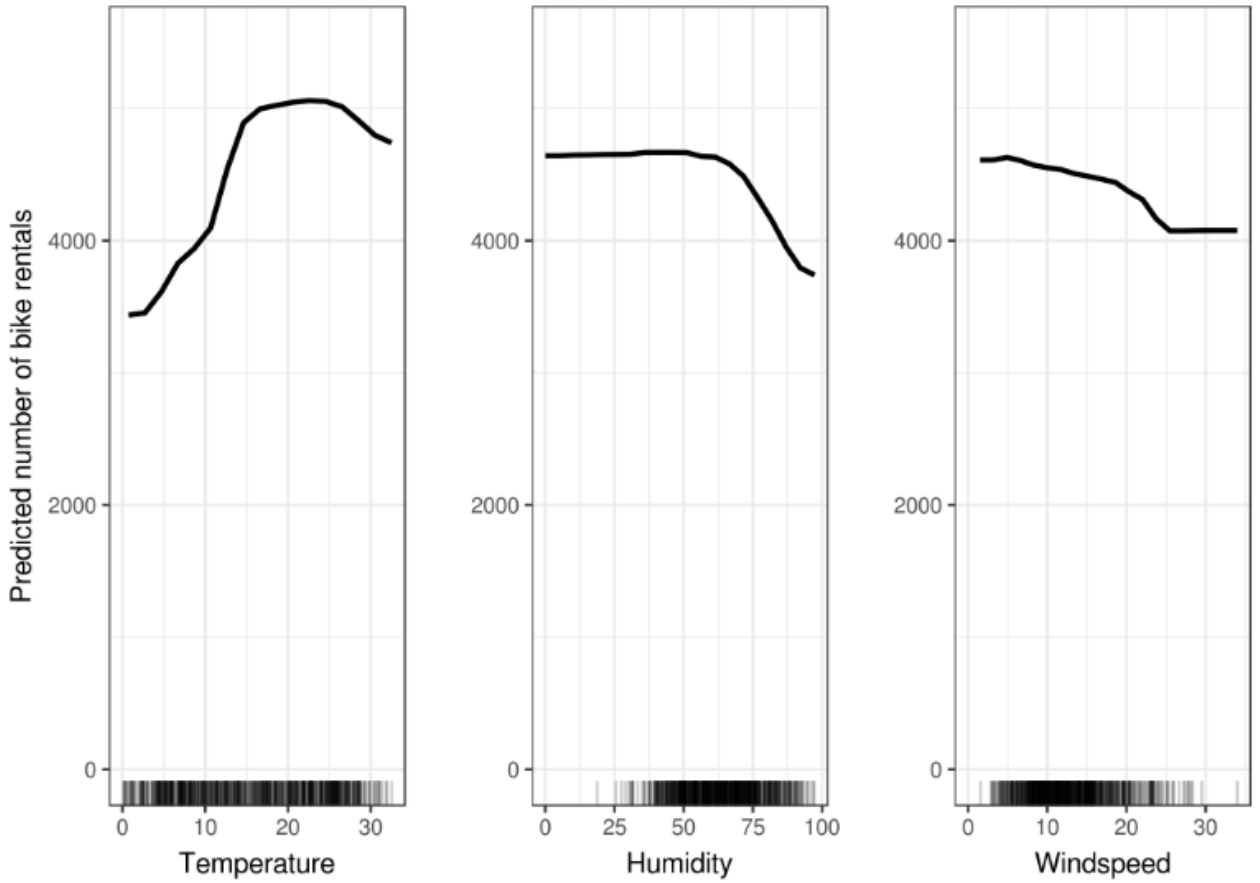
- Humans are the consumers of the explanations, ultimately.
- Here is where ‘Model Agnostic Methods’ sit and makes these models explainable
- Complex ‘Black Box Model learn with data to make predictions or find structures.
- Digitalize the ‘World’, like represent it as images, texts, tabular data and so on.
- This is the real world we observe

Method 1

Partial Dependence Plot (PDP)

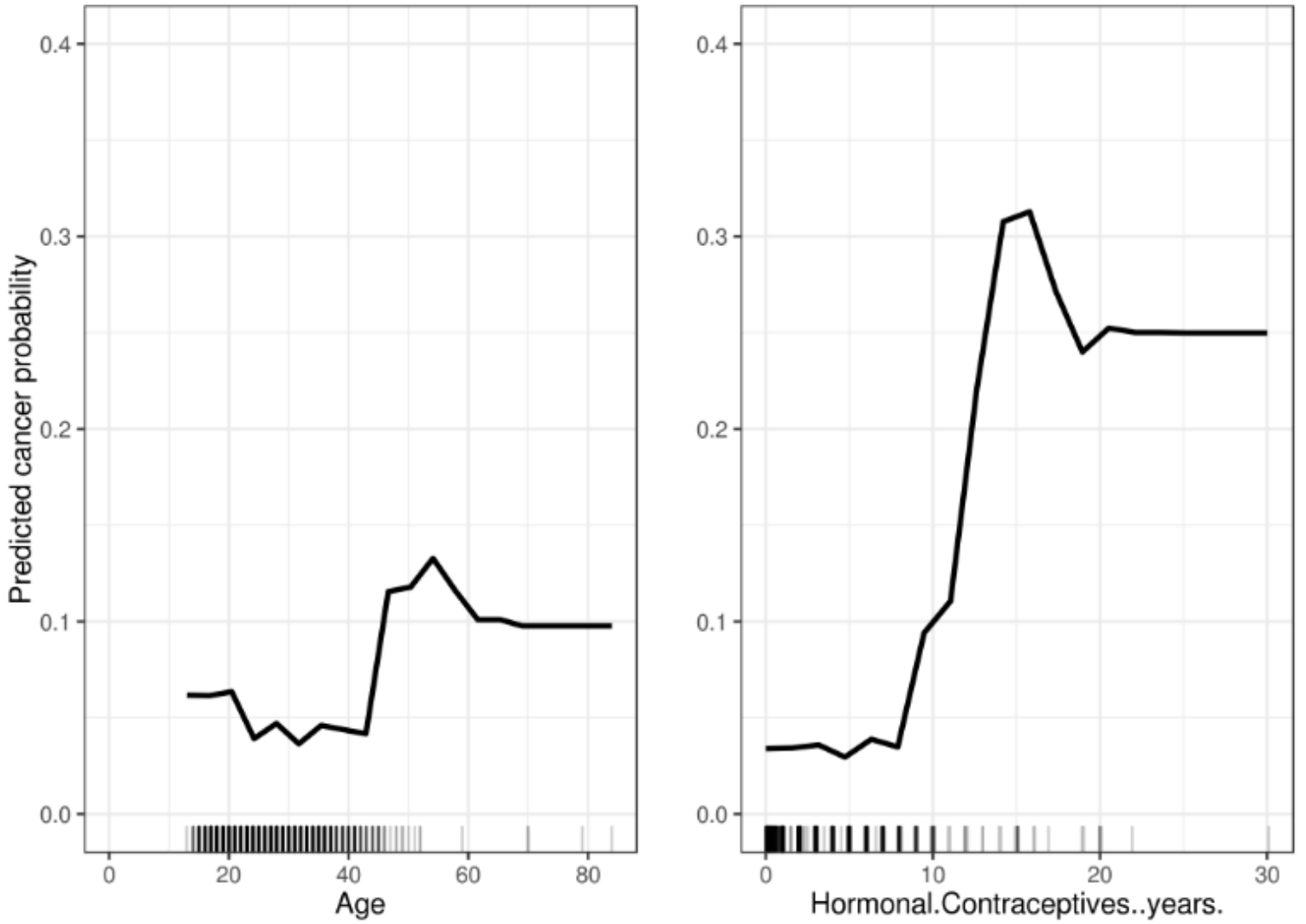
- The partial dependence plot shows the marginal effect of a feature on the predicted outcome of a previously fit model

On few features of bike rental dataset



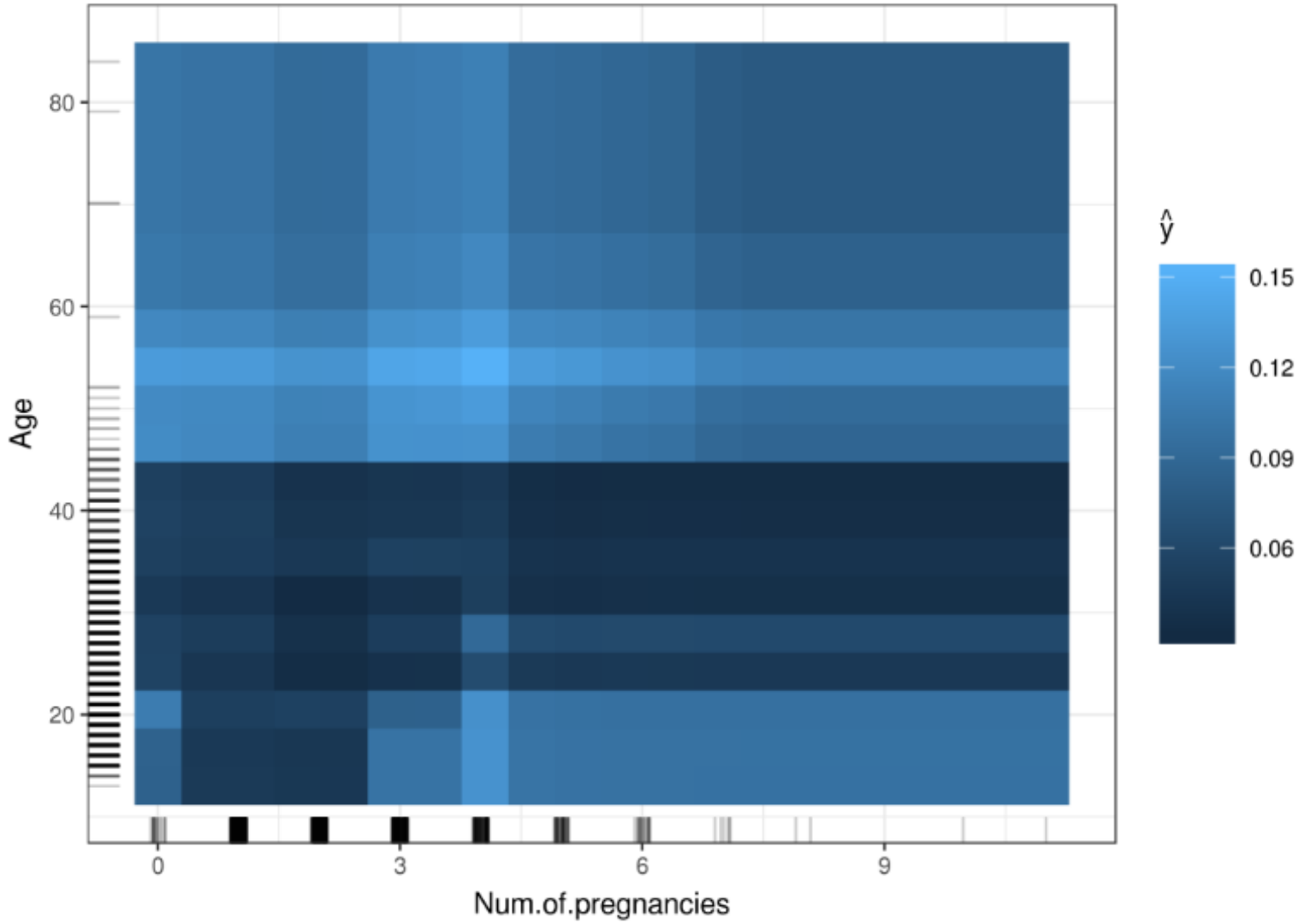
PDP Another dataset

On few features of cancer screening dataset



PDP
Two feature impact

On two features of cancer screening dataset



PDP Pros and Cons

Pros

- The computation is **intuitive**
- **Simple** to implement
- **Causal** interpretation

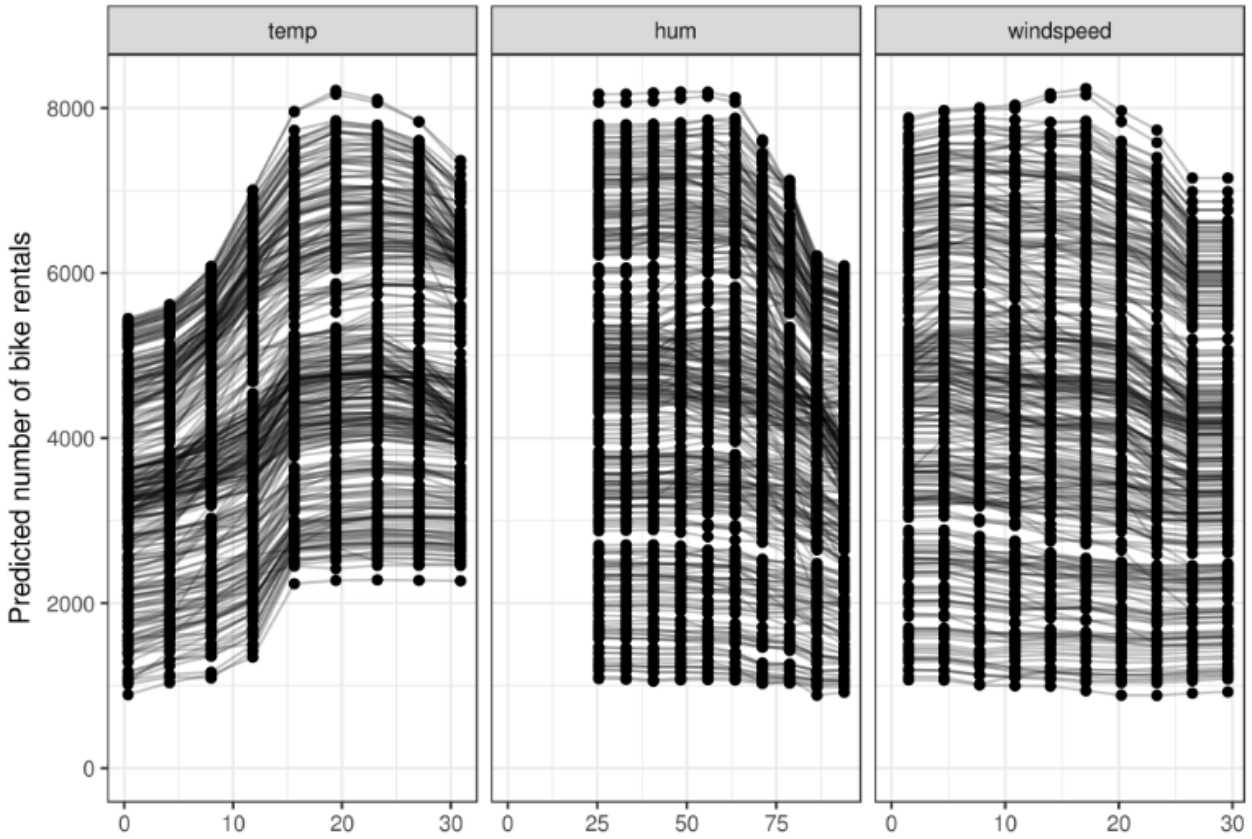
Cons

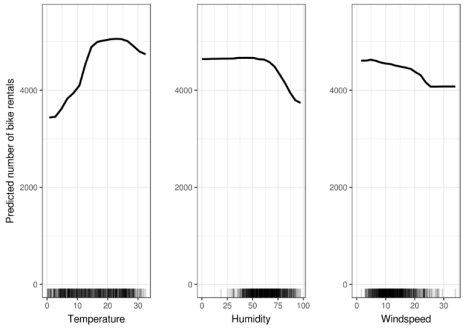
- The **maximum number of features** you can look at jointly is **two** or **three**.
- F**eature distribution** is important
- Assumption of **independence** is not true in real life datasets.
- **Heterogenous effects** might be hidden

Method 2
Individual Conditional Expectation (ICE)

- Plots one line per instance, representing how the instance's prediction changes when the feature changes.

On few features of bike rental dataset

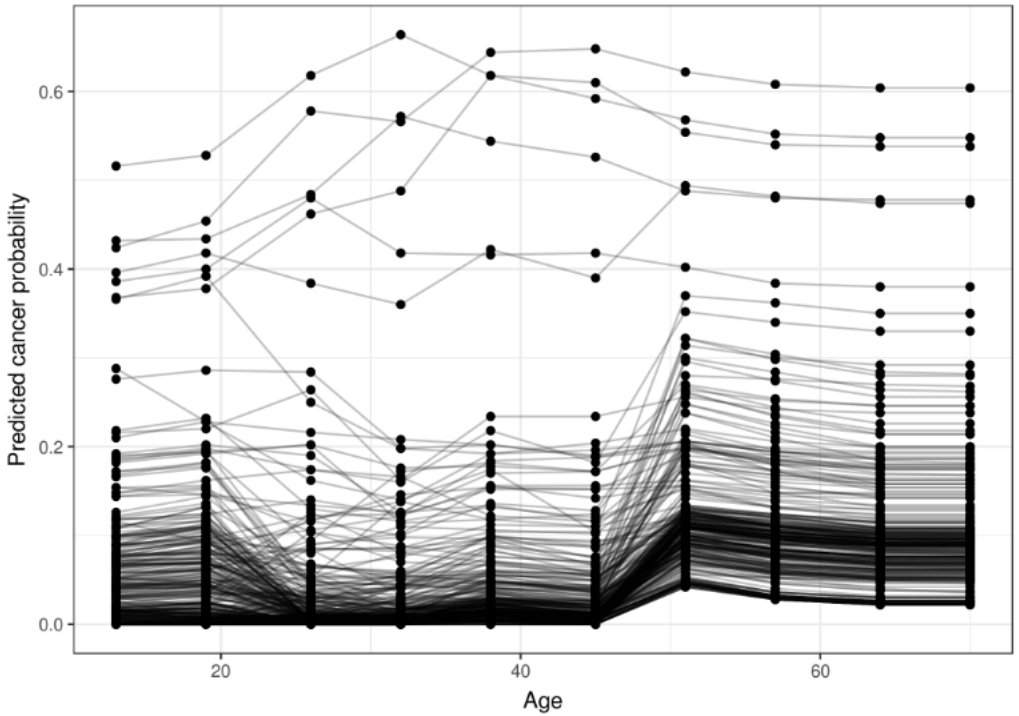




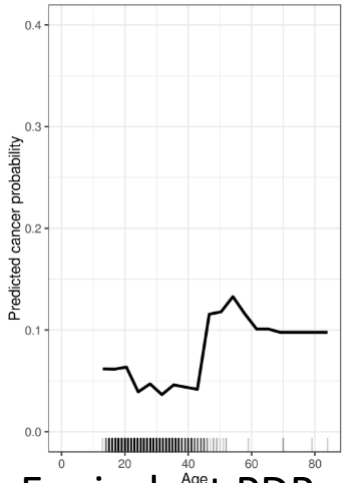
Equivalent PDP plot

ICE
Another dataset

On one feature of cancer screening dataset



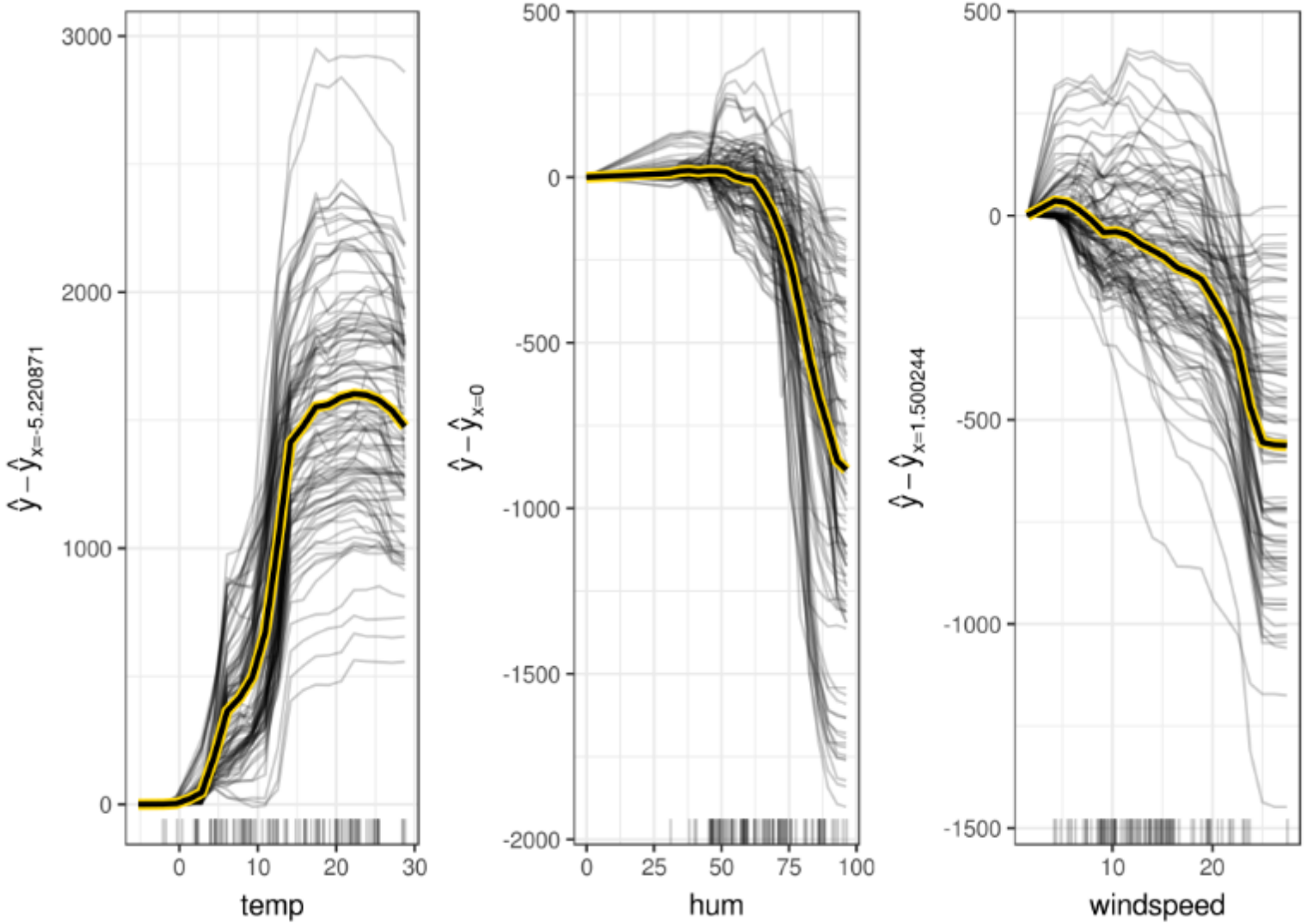
Drawback: Multiple points make is un-readable

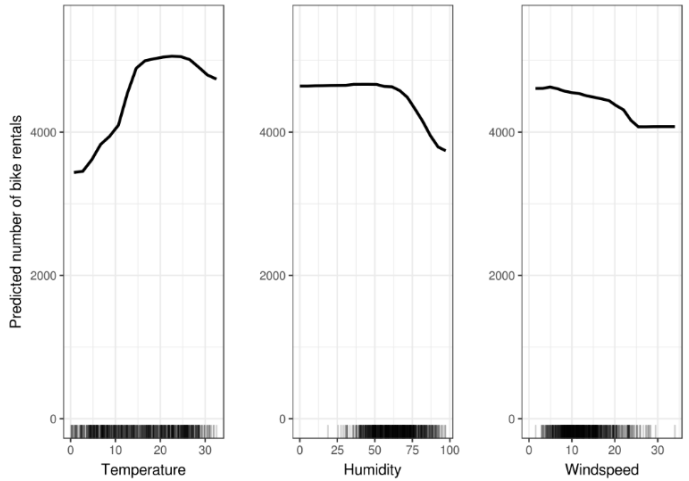


Equivalent PDP plot

Centered ICE Plot (c-ICE)

On few features of bike rental dataset

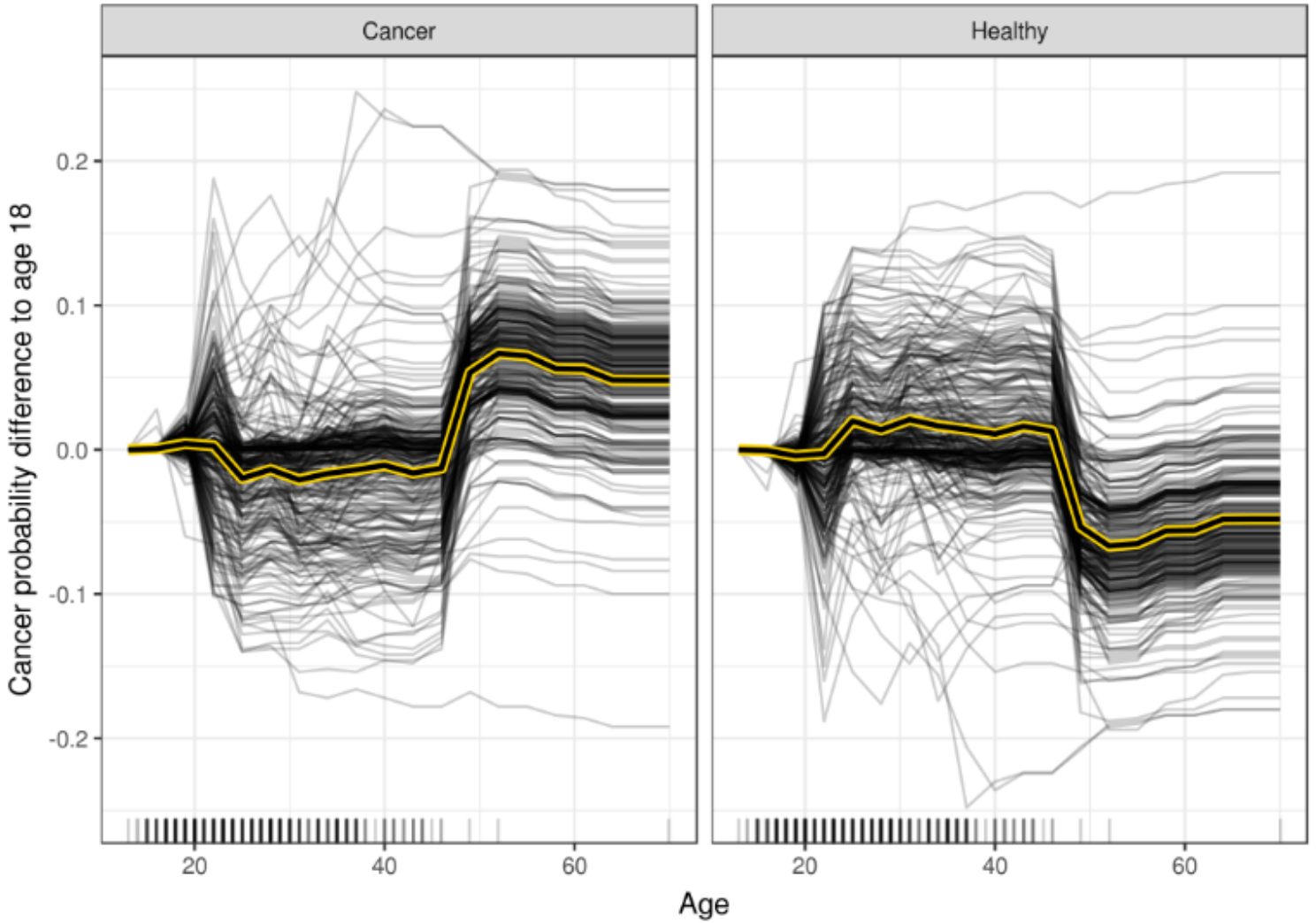


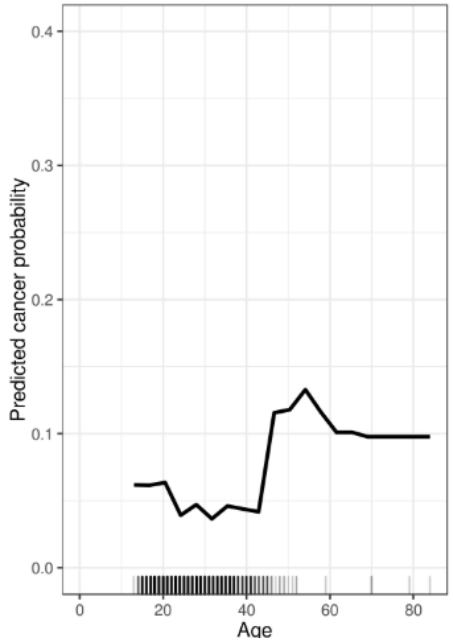


Equivalent PDP plot

c-ICE
Another dataset

On one feature of cancer screening dataset





Equivalent PDP plot

ICE Pros and Cons

Pros

- Even more **intuitive** to understand than PDP plots.
- Uncover **heterogeneous relationships**.

Cons

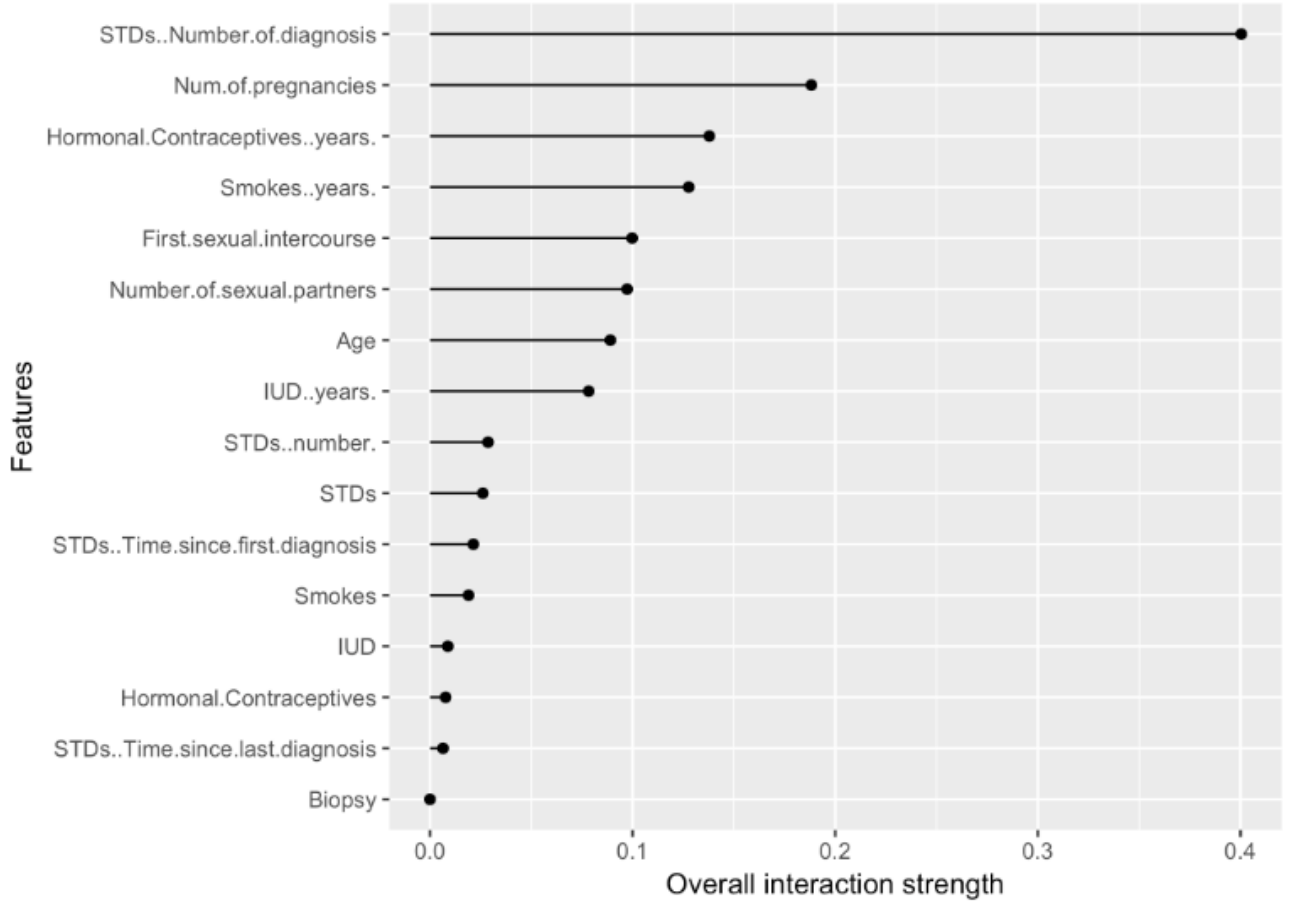
- ICE curves **can only display one feature** meaningfully
- Assumption of **independence** is not true in real life datasets.
- When many ICE curves are drawn the plot **can become overcrowded.**

Method 3
Feature Interaction Method

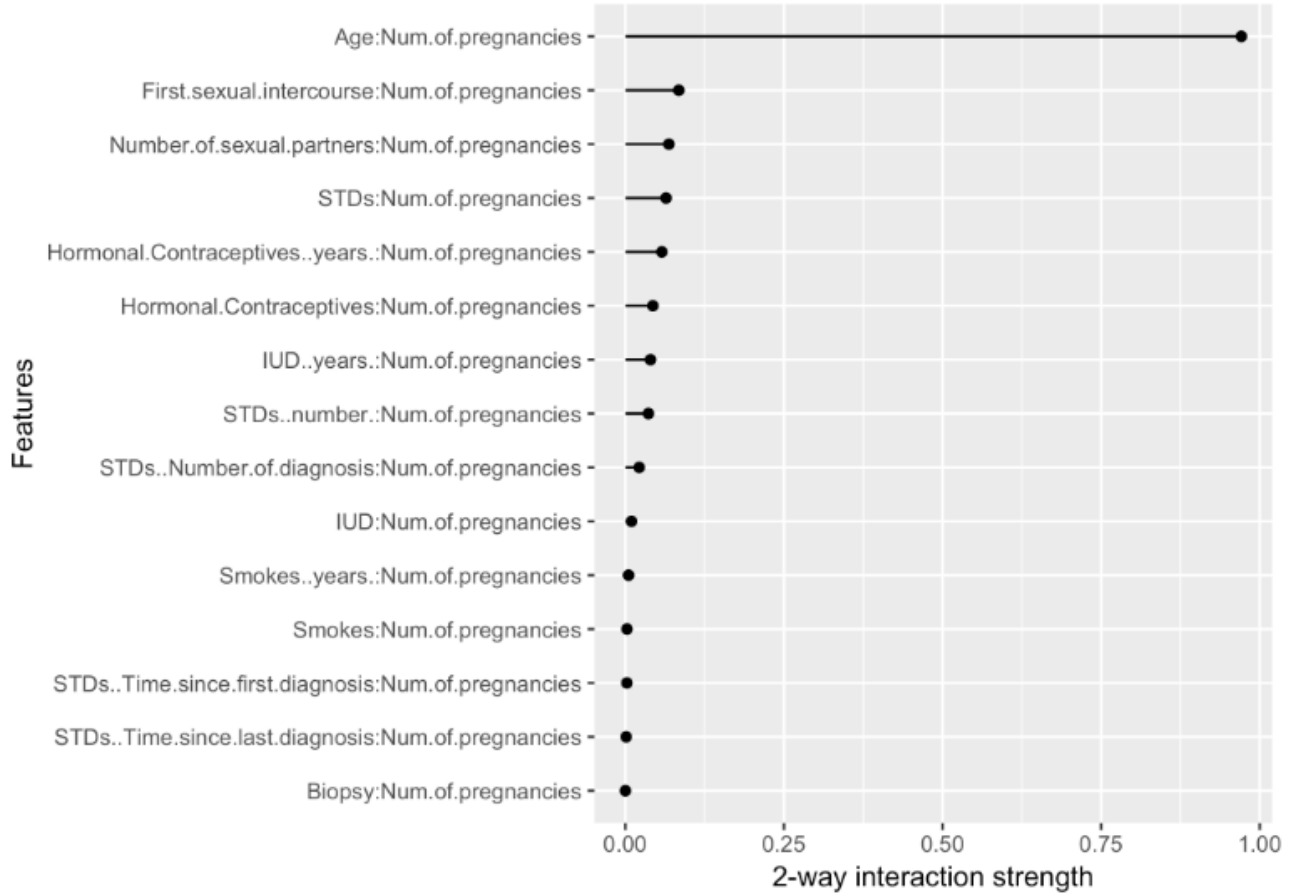
When features in a prediction model interact with each other, then the influence of the features on the prediction is not additive but more complex

Based on Friedman’s H-statistic

Feature Interaction – Examples (One with Other)



Two Way Feature Interaction



Feature Interaction Pros and Cons

Pros

- It can be explained using **PDP decomposition**.
- It has a **meaningful interpretation**
- It is **comparable** across features and models
- It **detects all kinds of interactions**, regardless of a specific form.
- Can explain **higher interactions,** i.e. > 3 features

Cons

- It's **computationally expensive**.
- From run to run and the results might **become unstable**
- Tells us the strength of interaction but **not the shape**
- Can't be applied on **image data**

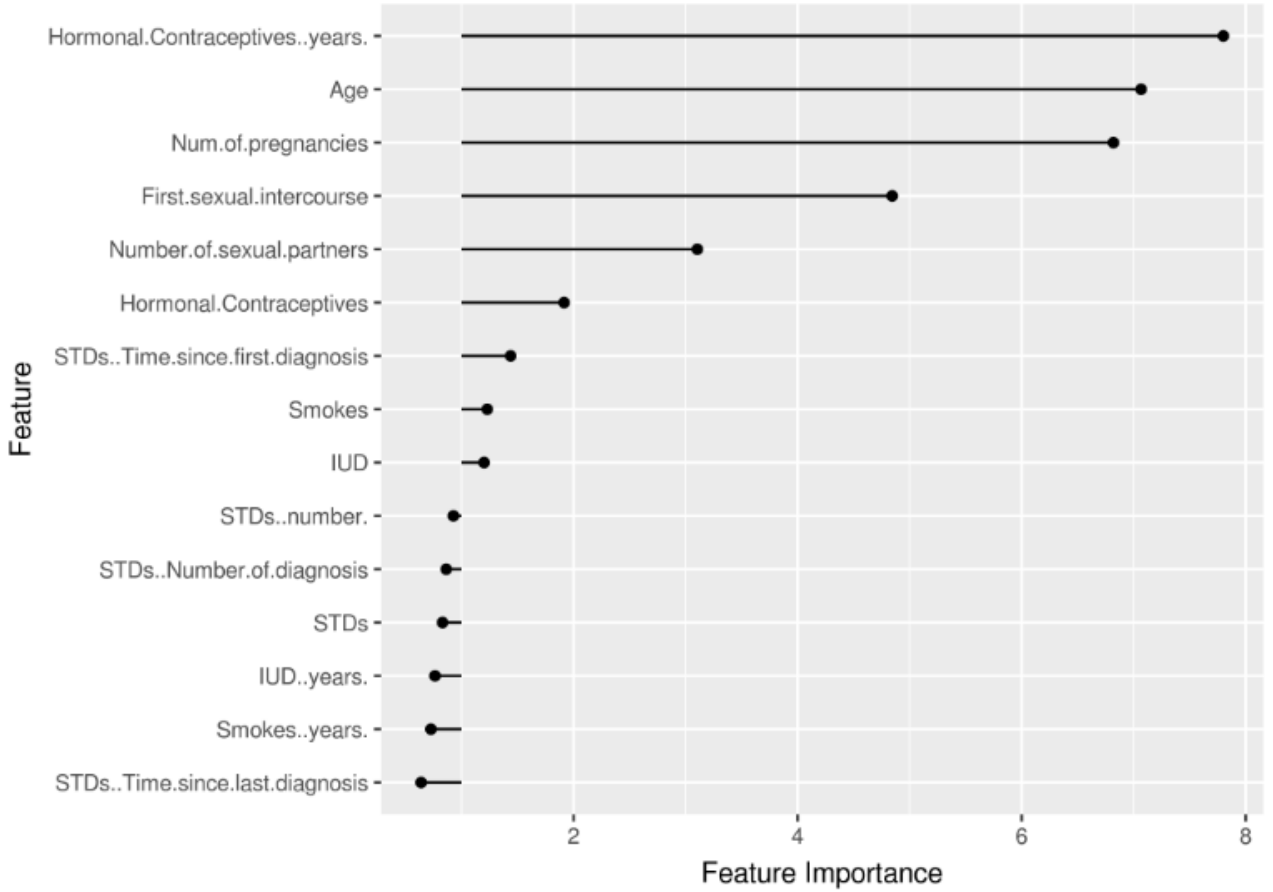
Method 4
Feature Importance Method

We measure a feature's importance by calculating the increase of the model's prediction error after permuting the feature. A feature is "important" if permuting its values increases the model error, because the model relied on the feature for the prediction.

Steps

1. Calculate the base error
2. For each feature
    i. Shuffle the feature to break relation
    ii. Calculate the error
    iii. Note the difference from base error
3. Sort error from max to min

Feature Importance Example



Feature Importance Pros and Cons

Pros

- Nice and **simple** interpretation.
- It provides a highly compressed, **global insight** into the model's behavior.
- Its measurements are **comparable** across different problems

Cons

- You need **access** to the **actual outcome target**.
- When **correlated**, the **permutation** feature importance measure can be **biased** by **unrealistic data instances**.
- Adding a **correlated feature** can **decrease** the **importance** of the **associated feature**.

Method 5
Global Surrogate Models

- A global surrogate model is an interpretable model that is trained to approximate the predictions of a black box model.
- We can draw conclusions about the black box model by interpreting the surrogate model.

Simply put

1. Take a black box model
    1. Get its underlying features and predictions

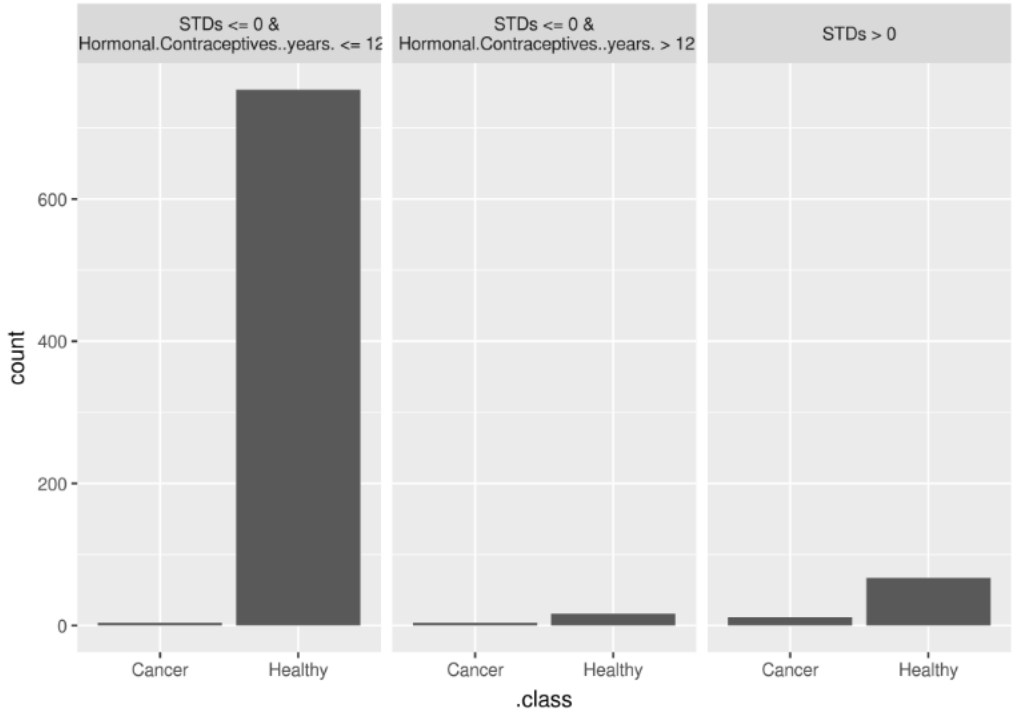
2. Train a simple explainable model on these features and prediction instead of actual target. Call it surrogate model

3. Measure how well surrogate model compares with the black box model by using Rsquared statistics

4. Use this surrogate model for explanations.

Global Surrogate - Example

- Based on a random forest based model predictions, we train a surrogate model using decision tree



Global Surrogate Model Pros and Cons

Pros

- The surrogate model method is very **flexible**.
- The approach is very **intuitive** and straightforward.
- Its measurements are **comparable** across different problems

Cons

- Can draw **conclusions about the model,** but not the data.
- Make a guess on **cut-off for R squared.**

Method 6
Local Surrogate Model Explainability (LIME)

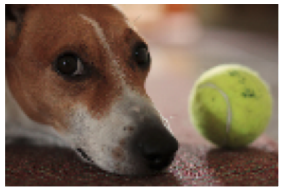
- Fitting local, interpretable models that can explain single predictions of any black-box machine learning model.
- LIME explanations are local surrogate models

Simply put

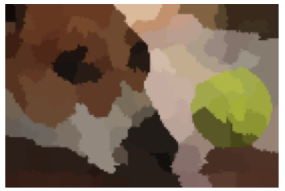
1. Take the specific result from the black box model that you want to explain
2. Generate/Get the data points around this point.
3. Train a simple surrogate model for these points
4. Measure how well surrogate model compares with the black box model by using Rsquared statistics
5. Use this surrogate model for explanations of the specific result.

https://arxiv.org/pdf/1602.04938.pdf

LIME for Image Data

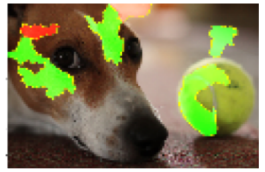


Segment the image

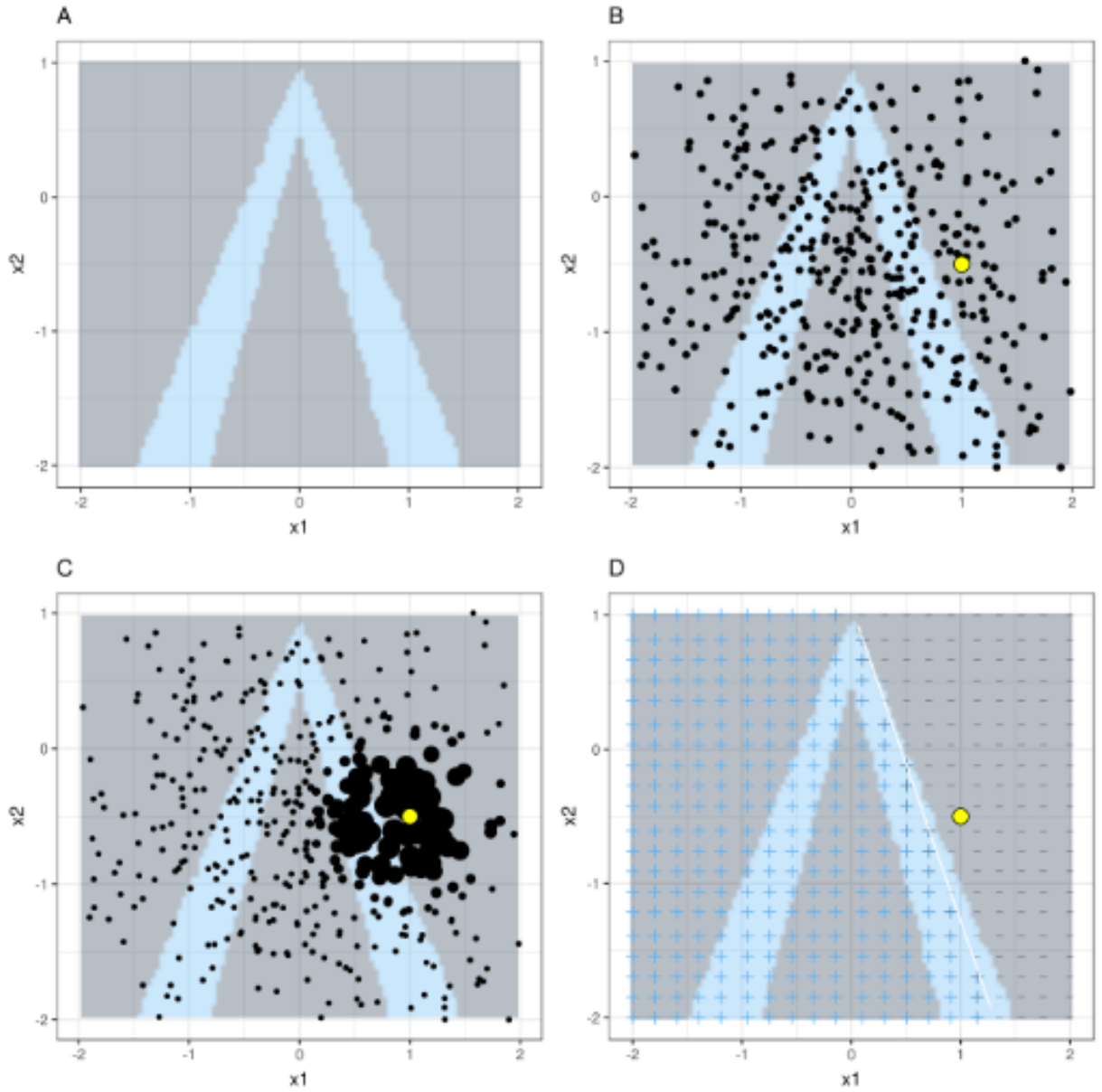


Creates several variants of the original image, each with a different combination of “fudged” super pixels and original pixels

- It computes its distance from the original image
- It uses the classifier to classify this variant.
- Create an optimal linear model of the “importance” of each image segment to the eventual class
- We can ask for the “top segment” or “top X segments” which explains the classification of this image as a tennis ball (corresponding to the highest coefficient in our linear model)



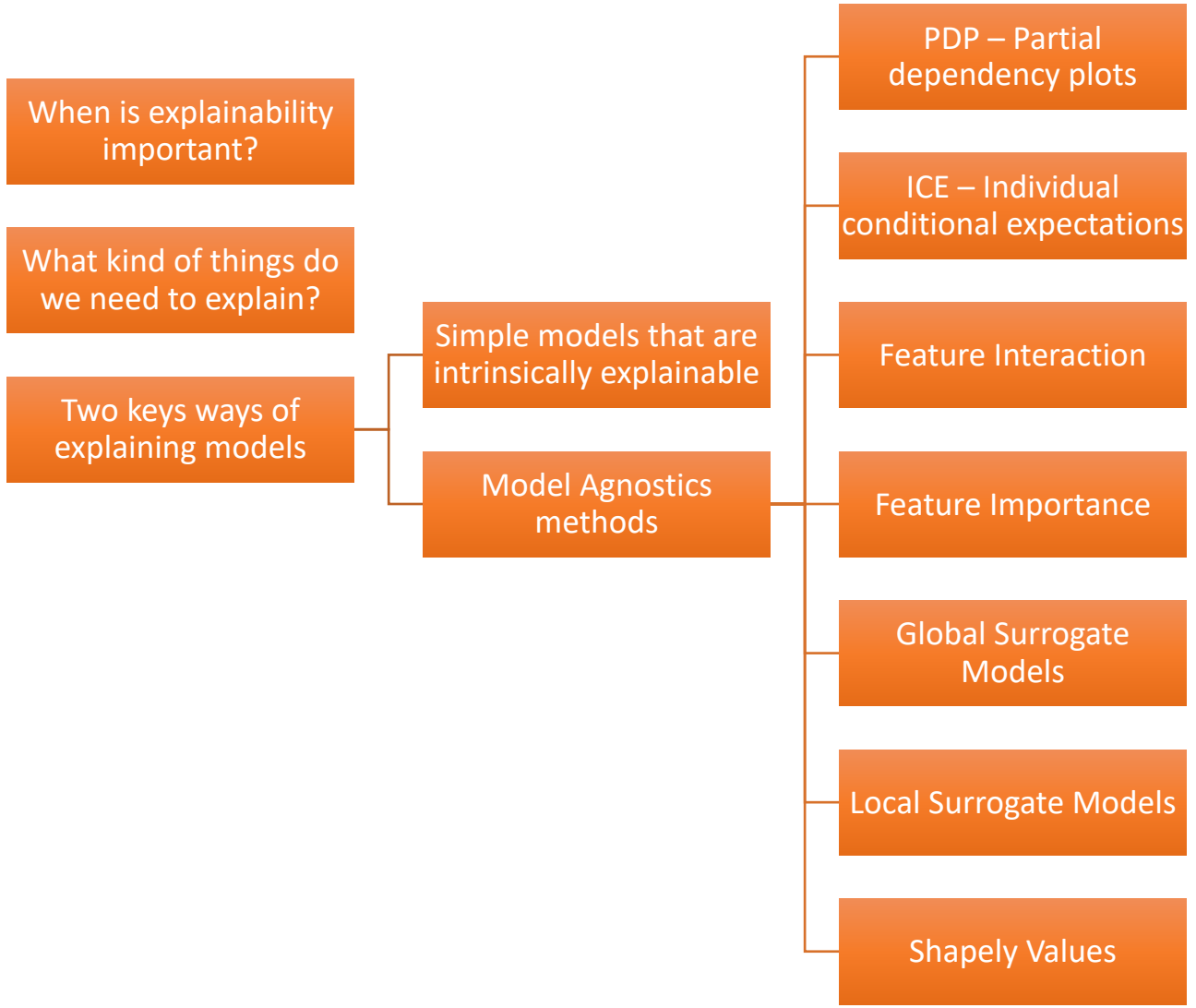
LIME for Tabular Data



Method 7
Shapley Value

- Predictions can be explained by assuming that each feature is a ‘player’ in a game where the prediction is the payout.
- The Shapley value - a method from coalitional game theory - tells us how to fairly distribute the ‘payout’ among the features.

Summary


When is explainability important?
What kind of things do we need to explain?
Two keys ways of explaining models
Simple models that are intrinsically explainable
Model Agnostics methods
PDP – Partial dependency plots
ICE – Individual conditional expectations
Feature Interaction
Feature Importance
Global Surrogate Models
Local Surrogate Models
Shapely Values

Questions