2695 Introduction to Machine Learning
Masters Program in Economics, Finance and Management

7a  Video Lecture

MODEL INTERPRETABILITY

What is interpretability and why is it important?

- The degree to which a human can understand the cause of a decision.
- Sometimes it is used interchangeably with explainability, sometimes a distinction is made
  - **Interpretable** Machine Learning focuses on designing models that are inherently interpretable.
  - **Explainable** Machine Learning tries to provide post hoc explanations for existing models.

- Why is it important
  - Social acceptance
  - Safety
  - Ethics
  - Scientific understanding
  - Regulation

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

Taxonomy of interpretability methods

- **Model-specific**

  Limited to specific model classes.
- **Model-agnostic**
  - Used on any machine learning model and are applied after the model has been trained (post hoc).

- **Local**

  Interpretation method explains an individual prediction.
- **Global**

  Interpretation method explains the entire model behavior.

Result of the interpretation method

- Feature summary statistic
- Feature summary visualization
- Model internals
- Data point
- Intrinsically interpretable model

Interpretable models

Interpretable models have transparent decision-making processes that allow users to trace how predictions are made

- Decision tree
- Linear regression
- Logistic regression

Interpreting linear regression

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n$$

- Interpretation of weight for numerical feature: An increase of feature x_i by one unit increases the prediction for y by Θ_i units when all other feature values remain fixed.
- Weight sign:
  - positive: as feature value increases, target value increases
  - negative: as feature value increases, target value decreases
- The importance of a feature increases with increasing absolute weight.
  - Features must be scaled for weights to be comparable.
- If features correlated: individual feature weights do not give the complete picture.

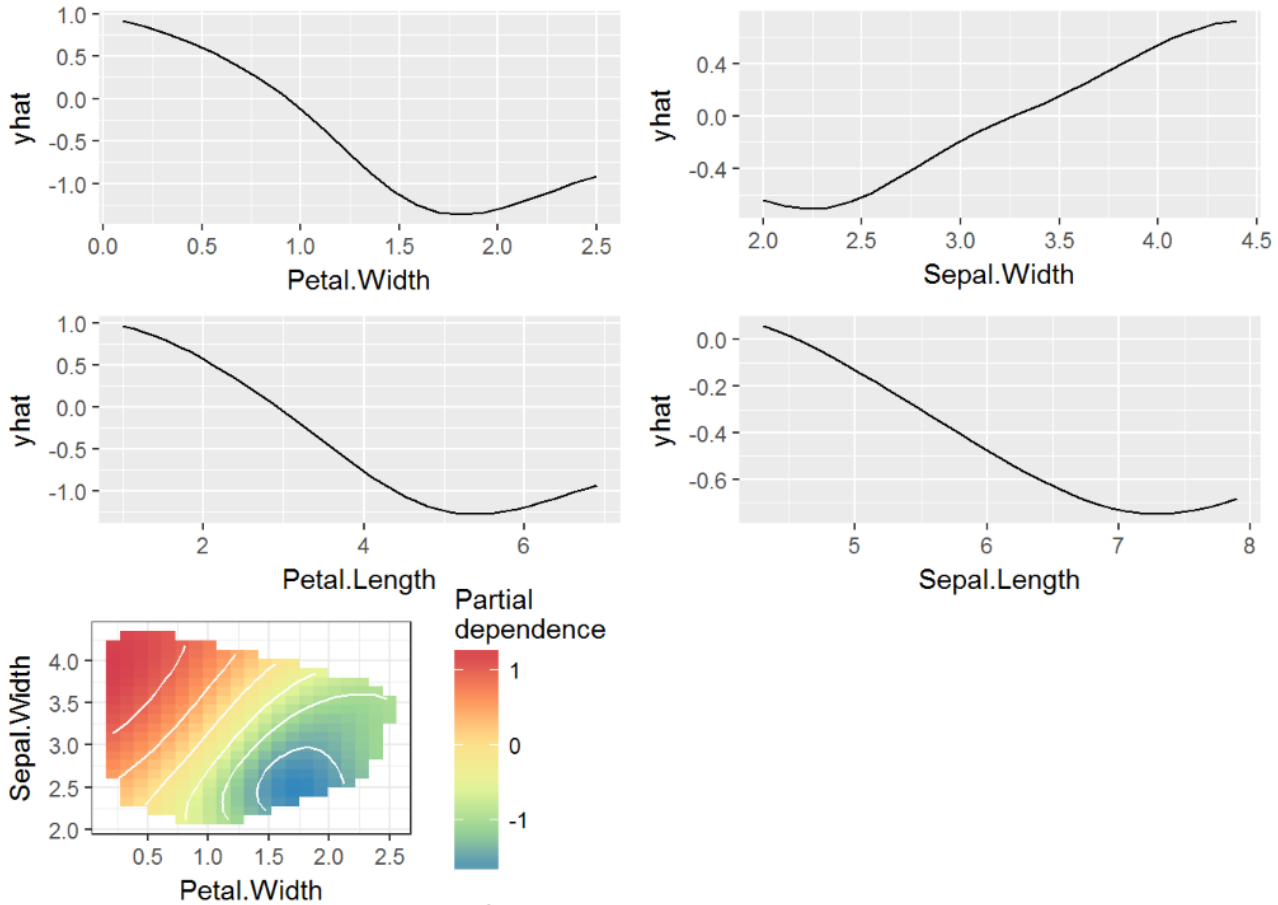
- Logistic regression: The coefficients show how each feature influences the probability of an outcome.

Model agnostic methods

- Partial Dependence Plot (PDP)
- Individual Conditional Expectation (ICE)
- Permutation Feature Importance
- Local Surrogate (LIME)
- SHAP (SHapley Additive exPlanations)

Partial dependence plot (PDP)

- It shows the marginal effect one or two features have on the predicted outcome of a machine learning model.
- It is a global method, considers all instances and shows the global relationship of a feature with the target.



x-axis: feature value

Depending on the implementation, y-axis can be:

- actual predicted value
- actual predicted probability
- change in the prediction metric relative to a baseline value
- custom metric

http://rstudio-pubs-static.s3.amazonaws.com/283647_c3ab1ccee95a403ebe3d276599a85ab8.html

How to calculate PDPs

1. Select feature (A)
2. Define grid of values (A1, A2, A3)
3. For each value of the grid:
   - Replace feature with grid value
   - Average predictions
4. Plot curve

A	B	C	Y
A1	B1	C1	Y1
A2	B2	C2	Y2
A3	B3	C3	Y3

A	B	C	Y	mean
A1	B1	C1	Y11	
A1	B2	C2	Y21	Y(A1)
A1	B3	C3	Y31	
A2	B1	C1	Y12	
A2	B2	C2	Y22	Y(A2)
A2	B3	C3	Y32	
A3	B1	C1	Y13	
A3	B2	C2	Y23	Y(A3)
A3	B3	C3	Y33	

X	A1	A2	A3
Y	Y(A1)	Y(A2)	Y(A3)

- Dataset with 3 features and 3 datapoints
- Feature A has three unique values: A1, A2, A3

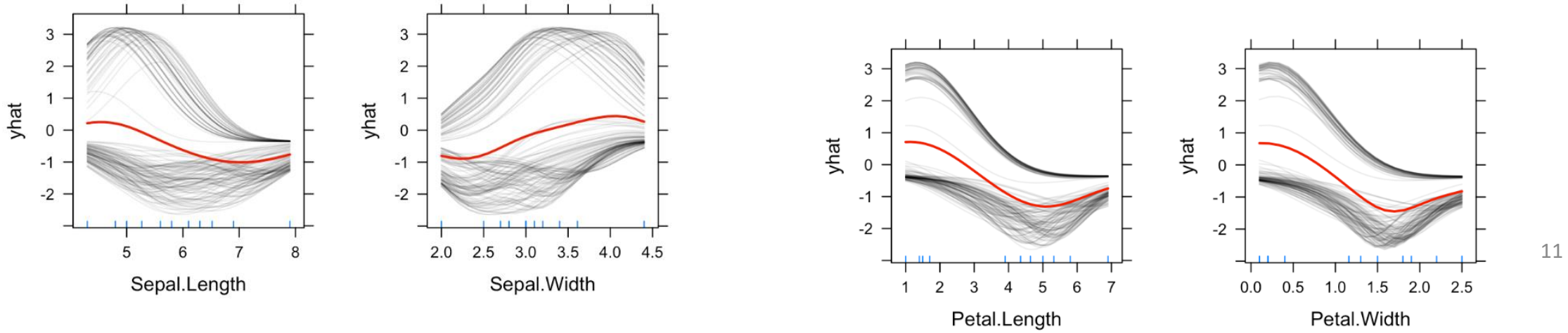
http://ethen8181.github.io/machine-learning/model_selection/partial_dependence/partial_dependence.html

Advantages and disadvantages of PDP

- Advantages:
  - Intuitive: The partial dependence function at a particular feature value represents the average prediction if we force all data points to assume that feature value.
  - If the feature for which we computed the PDP is not correlated with the other features, then the PDPs perfectly represent how the feature influences the prediction on average.
  - Easy to implement.
- Disadvantages:
  - Limited to two features.
  - The assumption of independence: assumes features for which the partial dependence is computed are not correlated with other features.
  - Heterogeneous effects might be hidden because PDP only show the average marginal effects.

Individual Conditional Expectation (ICE)

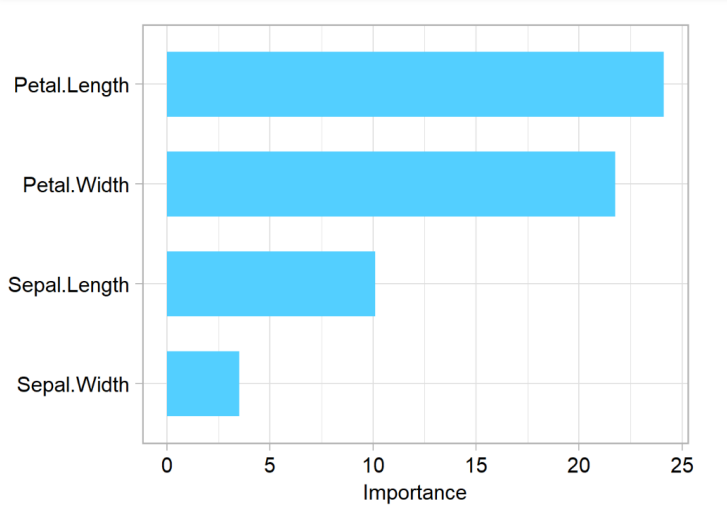
- The equivalent to a PDP for individual data instances is ICE plot.
- Individual Conditional Expectation (ICE) plots display one line per instance that shows how the instance's prediction changes when a feature changes.
- A PDP is the average of the lines of an ICE plot.
- The values for a line: computed by keeping all other features the same, creating variants of this instance by replacing the feature's value with values from a grid and making predictions with the black box model for these newly created instances.
- ICE curves suffer from the same problem as PDPs: If the feature of interest is correlated with the other features, then some points in the lines might be invalid data points according to the joint feature distribution.



Permutation Feature Importance

- Permutation feature importance measures the increase in the prediction error of the model after we permuted the feature's values.
- A feature is *important* if shuffling its values increases the model error- the model relied on the feature for the prediction.
- A feature is *unimportant* if shuffling its values leaves the model error unchanged- the model ignored the feature for the prediction.
- It could be applied on the training as well as the test set:
  - Training data: how much the model relies on each feature for making predictions.
  - Test data: how much the feature contributes to the performance of the model on unseen data.
- Different from Decision tree-based feature importance

  Gini index importance: for each feature go through all the splits for which the feature was used and measure how much it has reduced the Gini index compared to the parent node.



Advantages and Disadvantages of Permutation Feature Importance

Advantages

- Has nice interpretation: Feature importance is the increase in model error when the feature's information is destroyed.
- Provides a global insight into the model's behavior.
- Considers all interactions with other features. By permuting the feature, the interaction effects with other features are canceled.

Disadvantages

- Depends on shuffling the feature, which adds randomness to the measurement. When the permutation is repeated, the results might vary greatly.
- Decreases the importance of the correlated features by splitting the importance between both features.

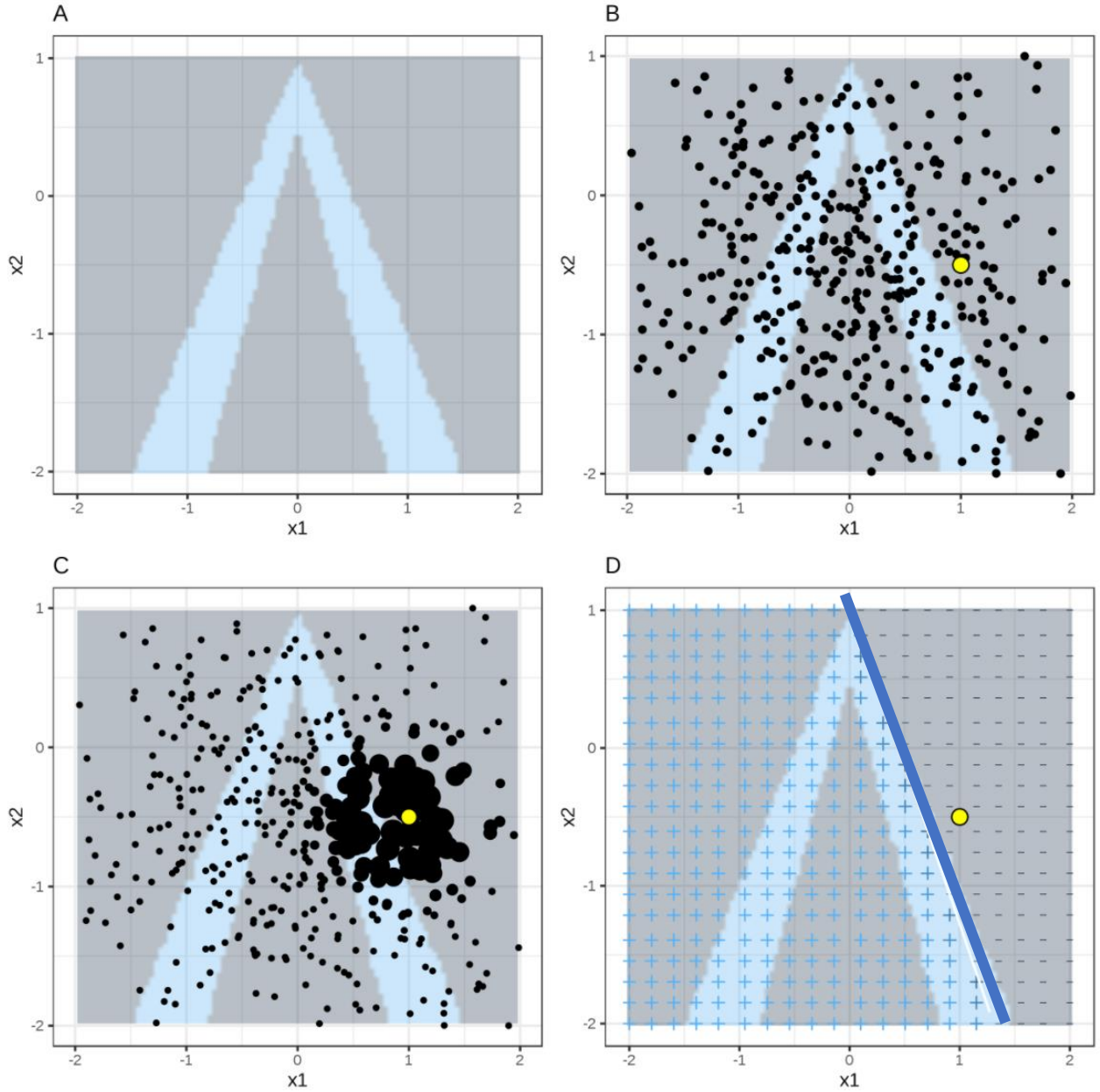
Local Interpretable Model-Agnostic Explanations (LIME)

- **Surrogate models:** trained to approximate the predictions of the underlying black box model.

- Basic idea:
  - LIME tests what happens to the predictions when we give variations of the data into the machine learning model.
  - LIME generates a new dataset consisting of perturbed samples and the corresponding predictions of the black box model.
  - LIME then trains an interpretable model on this new dataset, which is weighted by the proximity of the sampled instances to the instance of interest.

- **Local fidelity:** The learned model should be a good approximation of the machine learning model predictions locally, but it does not have to be a good global approximation.
- Interpretable model used: Ridge Regression, Lasso, Decision trees
  - Regression: linear model will predict the output of the black box model directly.
  - Classification: the linear model will predict the probability of the chosen class.

How LIME works for tabular data?



Advantages and disadvantages of LIME

Advantages

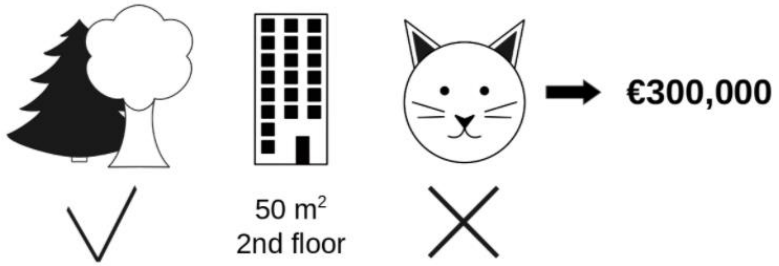
- When using Lasso or short trees, the resulting explanations are user friendly.
- LIME is one of the few methods that works for tabular data, text and images.
- LIME is very easy to use.

Disadvantages

- Choosing the correct neighborhood of a data instance depends on the parameters that should be tuned.
- Data points are sampled from a Gaussian distribution, ignoring the correlation between features. This can lead to unlikely data points which can then be used to learn local explanation models.
- The complexity of the explanation model must be defined in advance, it's a trade-off between fidelity and sparsity.
- Instability of the explanations: repeating the sampling process, might lead to different explanations.

Shapley values: illustrative example

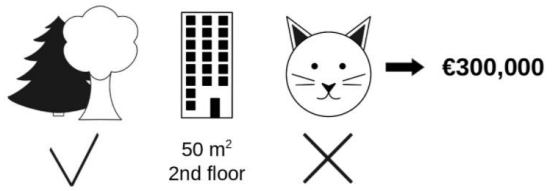
- We have trained a machine learning model to predict apartment prices.
- Predicts €300,000: apartment has an area of 50 m^2, is located on the 2nd floor, has a park nearby and cats are banned features: *area-50*, 2nd*floor, park-nearby, cat-banned*



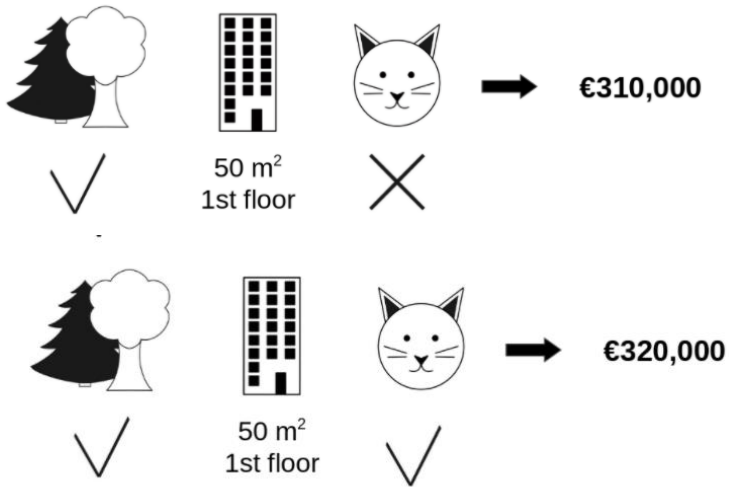
- The average prediction for all apartments is €310,000.

- Goal is to explain the difference between the actual prediction (€300,000) and the average prediction (€310,000).

Shapley values: illustrative example



our data instance



One sample repetition to estimate the contribution of *cat-banned* to the prediction when added to the coalition of *park-nearby* and *area-50*.

Shapley values: illustrative example

Coalitions needed for computing the exact Shapley value of the *cat-banned* feature value:

- No feature values
- *park-nearby*
- *area-50*
- *floor-2nd*
- *park-nearby + area-50*
- *park-nearby + floor-2nd*
- *area-50 + floor-2nd*
- *park-nearby + area-50 + floor-2nd*

- We replace the feature values of features that are not in a coalition with random feature values from the apartment dataset to get a prediction from the machine learning model.

- For each of these coalitions we compute the predicted apartment price with and without the feature value *cat-banned* and take the difference to get the marginal contribution.

- **The Shapley value is the (weighted) average of marginal contributions.**

SHAP (SHapley Additive exPlanations)

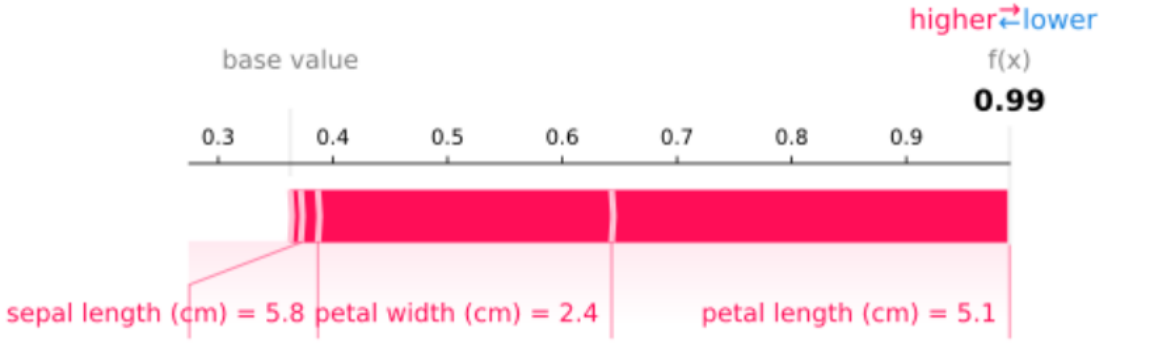
Advantages

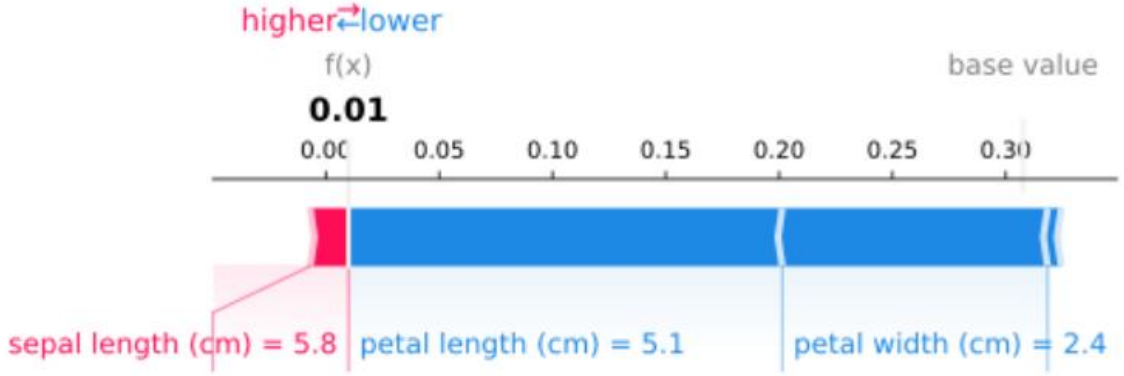
- The SHAP explanation method computes Shapley values.
- Global and Local interpretability
- With SHAP, global interpretations are consistent with the local explanations.
- Legally compliant method, because it is based on a solid theory and distributes the effects fairly.
- It has a fast implementation for tree-based models.

Disadvantages

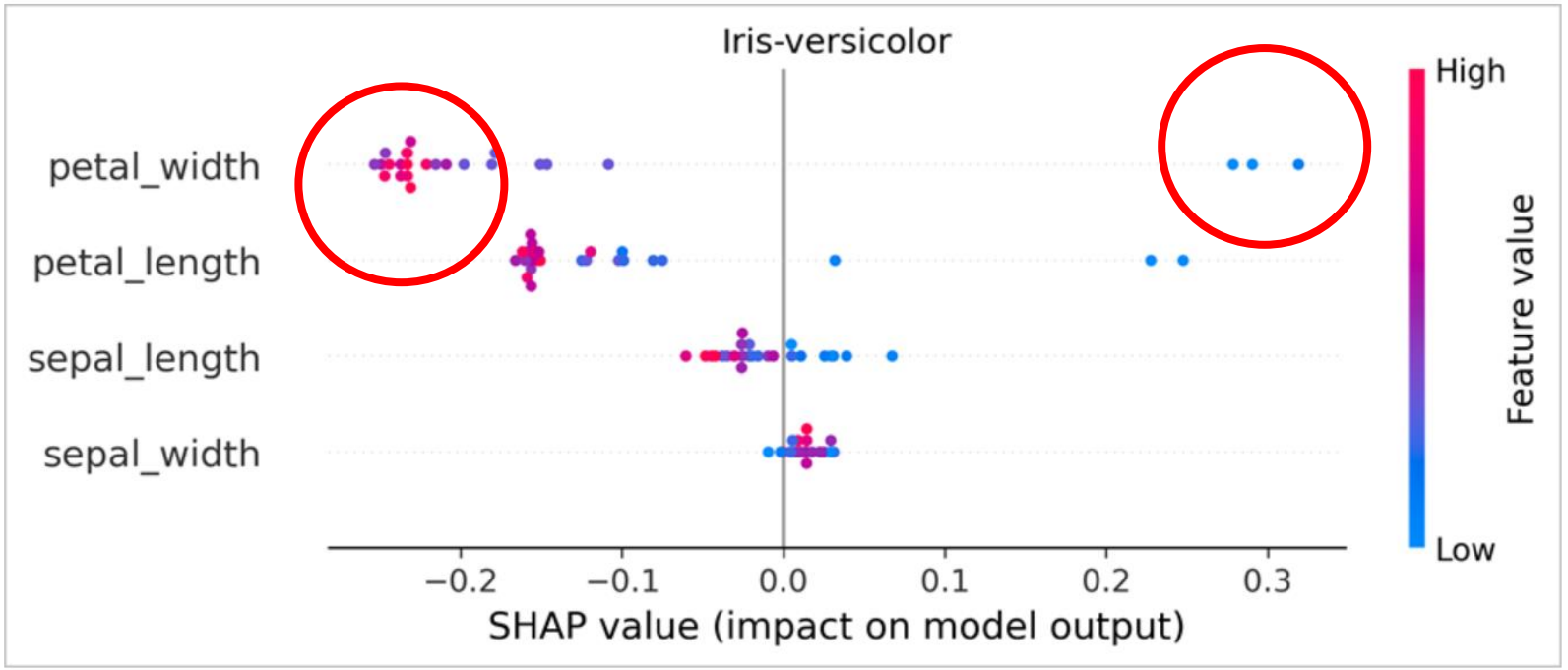
- To calculate the Shapley value: access to data is needed.
- Model-agnostic implementation KernalSHAP is slow. This makes KernelSHAP impractical to use when we want to compute Shapley values for many instances, unlike LIME.
- KernelSHAP ignores feature dependence.

SHAP Explaining a single prediction: Iris dataset



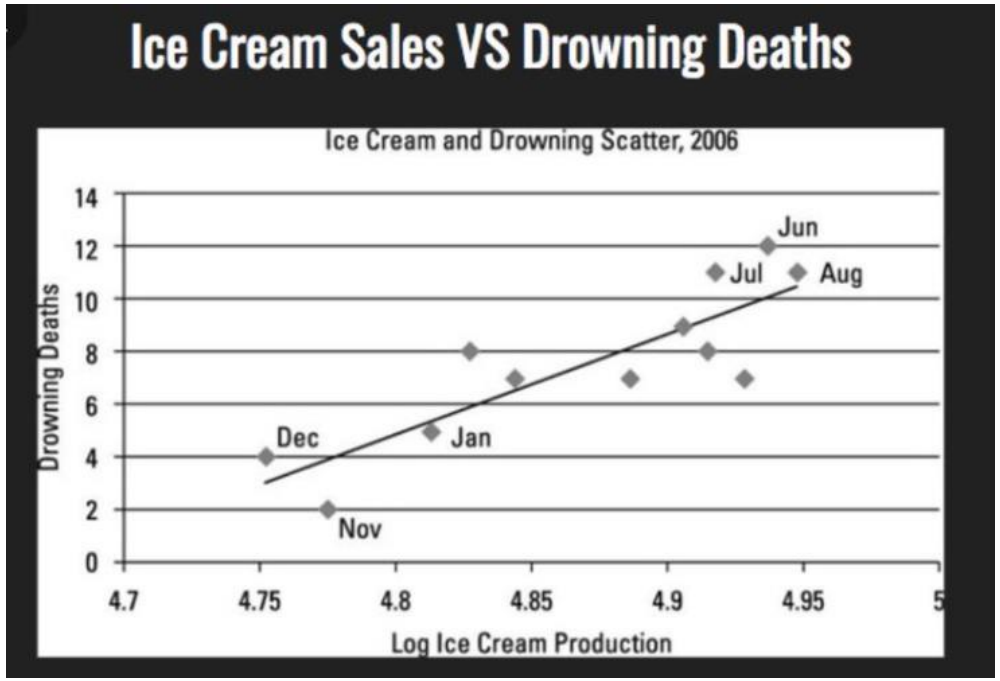


SHAP Global feature importance: Iris dataset



Correlation does not imply causation

- The target we are interested in cannot be directly observed through any single feature. Predicting most events, such as the likelihood a user will buy a given product, relies on different features, none of which directly measure the target.
- We should not assume that the patterns captured in models are causative, rather that they are correlations.
- Special analysis/tools/experiments should be used to confirm causation.



Spurious correlations [1]: mathematical relationship in which two or more events or variables are associated but not causally related, due to either coincidence or the presence of a certain third, unseen **confounding** factor.

[1] https://www.tylervigen.com/spurious-correlations

7b Video Lecture

FAIRNESS AND BIAS IN MACHINE LEARNING

Bias in ML

Systematic and repeatable errors in the model's predictions that create unfair outcomes, such as privileging one arbitrary group over others.

Different types of bias:

- Data bias
- Label bias
- Omitted variable bias
- Response bias
- Confirmation bias
- Survivorship bias
- …

Data bias

- Occurs when the training data does not accurately represent the real-world population, leading to systematic errors
- In a ML project: we need to check does that the training population match the population to which the model will be applied to.

Saturday Seminar | Rights | Mar 20, 2021

Facing Bias in Facial Recognition Technology

Brianna Rauenzahn, Jamison Chung, and Aaron Kaufman

In a National Institute of Standards and Technology report, researchers studied 189 facial recognition algorithms—"a majority of the industry." They found that most facial recognition algorithms exhibit bias. According to the researchers, facial recognition technologies falsely identified Black and Asian faces 10 to 100 times more often than they did white faces. The technologies also falsely identified women more than they did men—making Black women particularly vulnerable to algorithmic bias. Algorithms using U.S. law enforcement images falsely identified Native Americans more often than people from other demographics.

Omitted variable bias

- Occurs when one or more important variables are left out of the model.
- Predicting customer churn
  - Omitted variable: appearance of a competitor
- Model used to predict the probability of death for patients with pneumonia so that high-risk patients could be admitted to the hospital while low-risk patients are treated as outpatients.
  - Model learned: pneumonia patients who were asthmatic had lower risk of dying from pneumonia compared to patients who were not asthmatic.
  - Omitted variable: patients with a history of asthma who exhibited symptoms of pneumonia usually were admitted not only to the hospital but directly to the Intensive Care Unit, more effective care.
  - Researchers discovered this by analyzing interpretable model

Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission

Response (activity) bias

This bias refers to the disproportionate oversampling from an unrepresentative subpopulation:

- online reviews are biased toward extremes
- social media trends reflect the loudest voices
- E-commerce data favoring frequent shoppers
- health & fitness apps only reflect motivated users

Historical bias

- Bias of the world: models inherit and reinforce past inequalities, stereotypes, or discriminatory patterns
- Data has unwanted properties that are regarded as biased given a perfect sampling and feature selection.
- Replicates decisions made in the world that were biased.
- Amazon:
  - Launched a system for automatically screening job applicants.
  - Trained on 10 years worth of job applications and their outcomes.
  - Historical data: most employees at Amazon were male particularly in technical roles.
  - Algorithm learned that men were more suitable candidates.

Amazon scraps secret AI recruiting tool that showed bias against women

Fairness

- No universal definition of fairness exists, seems to depend on different preferences and outlooks, and varies by culture and context.
- Fairness is the absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics.
- Fairness is lack of bias.
- A given algorithm is said to be **fair**, or to have **fairness** if its results are independent of some features, we consider to be sensitive (ethnicity, age, family status,…).

A **protected attribute** is a feature along which bias can occur.

A **protected group** is a vulnerable population that has a specific value of the protected attribute(s).

Bias audit

- Dataset audit
- Pre-processing audit
- Post-processing audit
  - Naïve approach:
    - Apply methods for **interpretability** to understand which features are most important to the model.
    - Limitations:
      - Bias may actually exist in features correlated with the protected feature (example: zip code).
      - Just because a protected feature is important, does not mean the model is biased.
      - Even if the protected features are not important, bias still might be present.

Confusion matrix: loan default

Predicted Label \ True Label	1	0
1	**True Positive** Model indicates a person *will* not pay back the loan, and that person *will* not pay back the loan	**False Positive** The model indicates a person *will* not pay back the loan, and that person *will* pay back the loan
0	**False Negative** Model indicates a person *will* pay back the loan, but that person *will* not pay back the loan	**True Negative** Model indicates a person *will* pay back the loan, and that person *will* pay back the loan

Additional metrics

Calculate across different groups:

- low-income vs high income neighborhoods
- young vs old

fall-out or false positive rate (FPR)

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{N}} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} = 1 - \mathrm{TNR}$$

miss rate or false negative rate (FNR)

$$\mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{P}} = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TP}} = 1 - \mathrm{TPR}$$

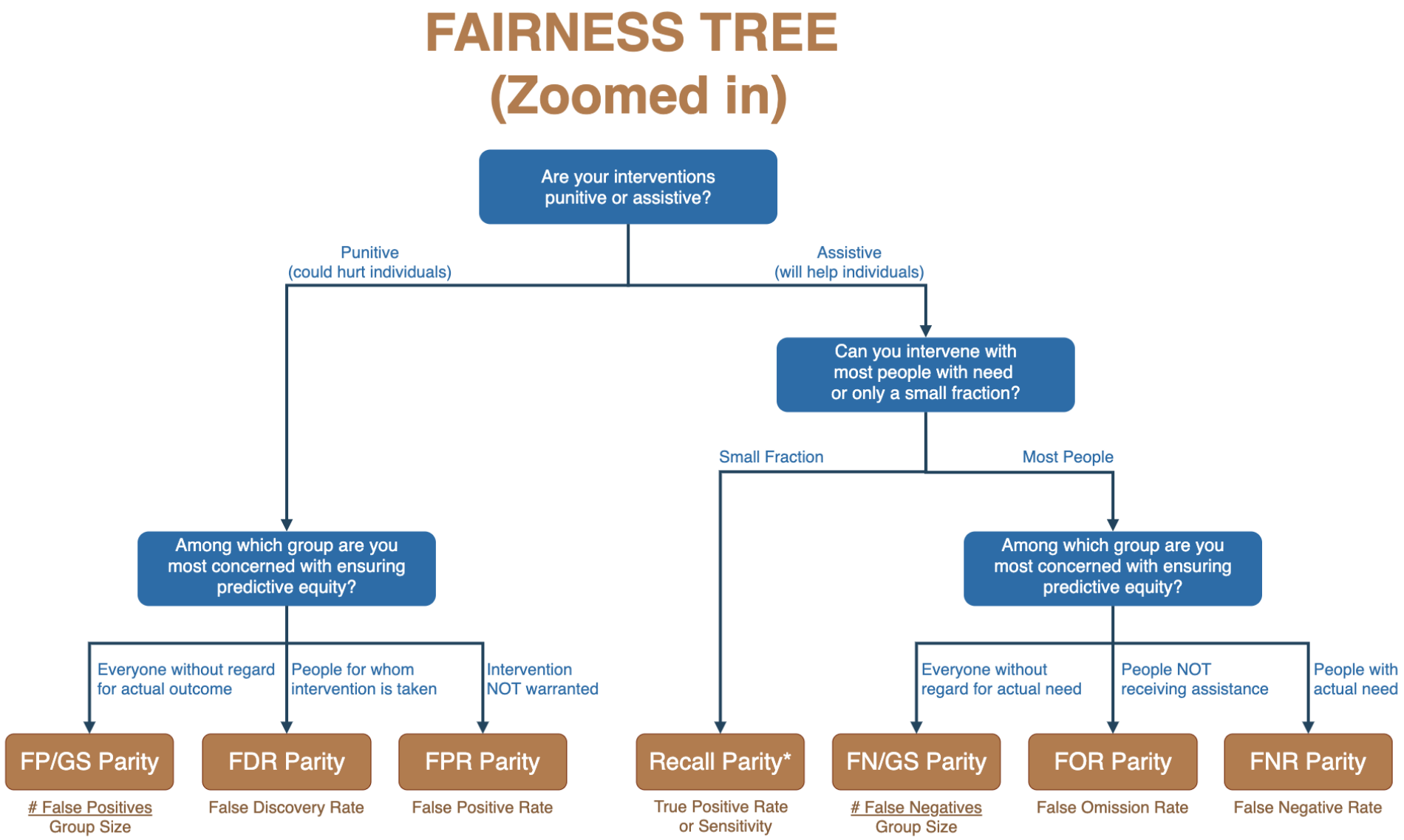
false discovery rate (FDR

$$\mathrm{FDR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TP}}$$

false omission rate (FOR)

$$\mathrm{FOR} = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TN}}$$

Fairness Tree



http://www.datasciencepublicpolicy.org/projects/aequitas/

Parity measures

- **Demographic parity:** equal proportion of positive predictions in each group:
  - (Loan rejection) Same proportion of loans rejected in each group
  - (University admissions) Same admission rate among each group

- **Equal opportunity:** equal True Positive Rate (Recall) across all groups.
- **Equality of Odds**: equal True Positive Rate (Recall) and False Positive Rate (FPR) across all groups.
- …

sensitivity, recall, hit rate, or true positive rate (TPR)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} = 1 - \mathrm{FNR}$$

fall-out or false positive rate (FPR)

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{N}} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} = 1 - \mathrm{TNR}$$

Accounting for bias

Preprocessing methods

Focused on transforming the dataset.

- Use oversampling and undersampling to decrease sampling bias in the dataset.
- Have group of people with diverse demographics labelling the dataset.
- Ensure that protected attributes do not influence the output of a machine learning model (naïve approach remove them).

In-processing methods

Focused on adjusting the machine learning algorithm.

- Include fairness constraints.
- Include evaluation metrics other than accuracy (FPR, FNR, FDR, FOR across different groups).

Postprocessing methods

Focused on altering the model's internals and predictions after the model has been trained.

- Tune thresholds to decrease false positives or false negatives.
- Calibrate probabilities of class membership so that the probabilities are closer to the true likelihood.
- Ensure that the model’s outputs can be explained and understood.

Fairness and bias

- Fairness is context dependent.
- All datasets are biased, and all models are unfair.
- Accounting for bias will create tradeoffs: impossible to satisfy all existing fairness criteria at the same time