Data/AI Projects Methodology

Chapter 2 : Open Challenges
Focus on AI Ethics & Explainability

Medina HADJEM

2022

Introduction

- **Data/IA projects still face many challenges :**
- **Functional/ Business Challenges**
  - Solve the right problem
  - Don’t re-invent the wheel
  - Solve the problems on time
  - Ensure the real usage
  - Take change & people into account

- ***We don’t fail because of the math… we fail because we don’t understand how people will use the math***

Introduction

- **Data/IA projects still face many challenges :**

- **Data Challenges**

  - Data Knowledge
  - Data Governance
  - Data Availability
  - Data Accessibility
  - Data Usability
  - Data Quality
  - Data Anonymization

Introduction

- **Data/IA projects still face many challenges :**

- **Technical Challenges**

  - AI projects Industrialization
  - AI projects monitoring
  - Models monitoring & backtesting
  - IT integration
  - Adapted technical environment

Introduction

- **Data/IA projects still face many challenges :**

- **Ethical Challenges**

  - Laws and regulations
  - Ethical accountability
  - Legal rights of an individual
  - Individuals' privacy and anonymity
  - Data ethical availability for its intended use
  - Data validity for its intended use
  - Data/model bias
  - Transparency need and achievement
  - Results interpretation => Explainable AI (XAI)

AI Ethics Challenges : Overview

When AI goes awry …

- **Deepfakes** : AI generated videos, audios, texts with intent to deceive
- **Bias AI** : regarding gender, race, and age biases (Ex : recruitment)
- **Facial Recognition Errors** : Ex : misidentification of a person as a criminal

Why AI Ethics ?

- Potential Harms caused by AI Systems

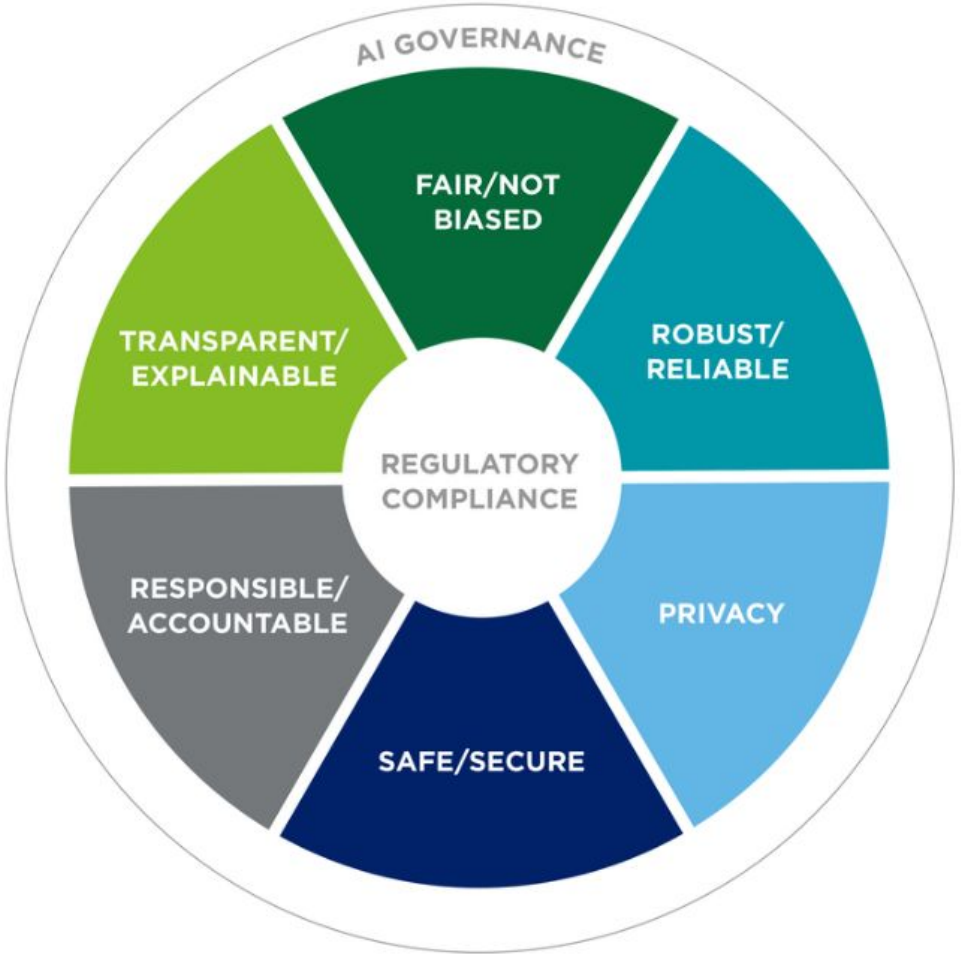
  - Bias & Discrimination
  - Non-transparent, Unexplainable, or Unjustifiable Outcomes
  - Invasions of Privacy :
  - Isolation and Disintegration of Social Connection
  - Denial of Individual Autonomy, Recourse, and Rights
  - Unreliable, Unsafe, or Poor-Quality Outcomes

What is AI Ethics ?

"AI ethics is a set of values, principles, and techniques that employ widely accepted standards of right and wrong to guide moral conduct in the development and use of AI technologies." [1]

1 : Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute. https://doi.org/10.5281/zenodo.3240529

AI Ethics Framework



source : https://www.forbes.com/sites/insights-ibmai/2020/03/26/trust-at-the-center-building-an-ethical-ai-framework/?sh=235835027bc7

Existing Initiatives

- The Institute for Ethical AI & Machine Learning (UK Research Center)
- Berkman Klein Center (Harvard based hub)
- European AI Alliance

  https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

- The Open Robo Ethics Institute (ORI) : non profit think-tank
- UK Government : Data Ethics Framework : https://www.gov.uk/government/publications/data-ethics-framework
- Responsible Computer Science Challenge (RSSC)

Existing technical tools

Technical Resources for Exploring Fairness Tools:

- University of Chicago's open source bias audit toolkit for machine learning developers

  https://dsapp.uchicago.edu/projects/aequitas/

- Datasets and software for detecting algorithmic discrimination from TU Berlin and Eurecat)

  http://www.fairness-measures.org/

  https://github.com/megantosh/fairness_measures_code/

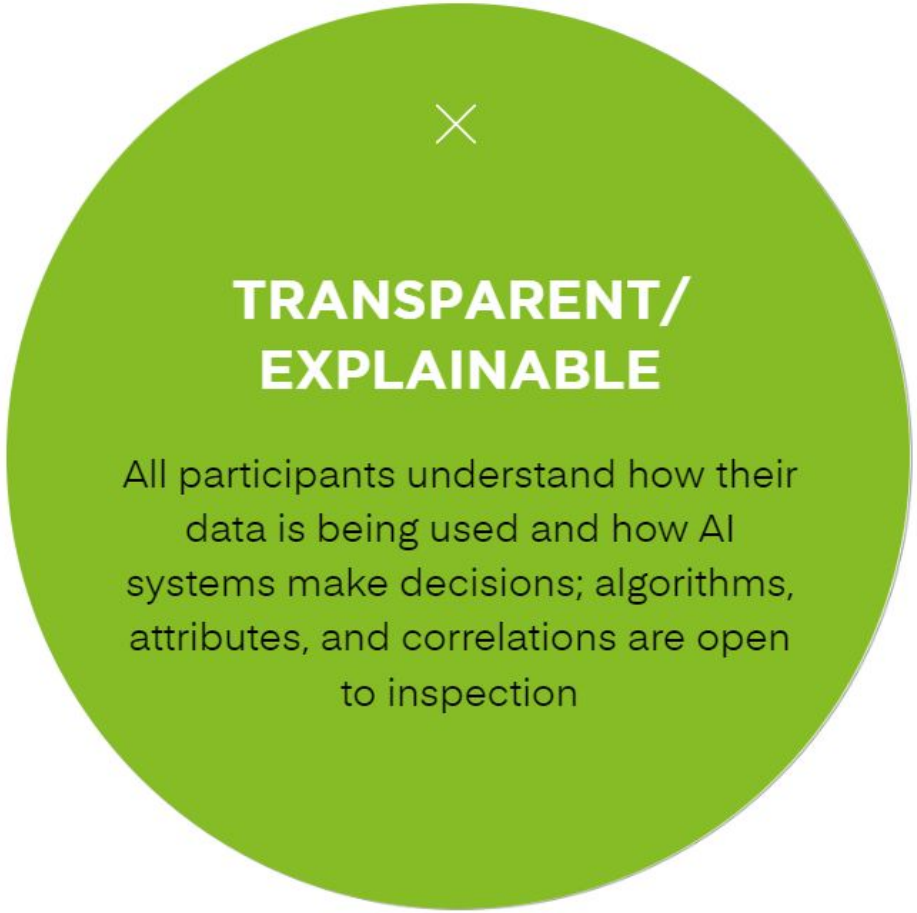
- Fairtest unwarranted association discovery platform from Columbia University

  https://github.com/columbia/fairtest

- IBM's Fairness 360 open source toolkit

http://aif360.mybluemix.net/#

Focus on Transparency/Explainability



source : https://www.forbes.com/sites/insights-ibmai/2020/03/26/trust-at-the-center-building-an-ethical-ai-framework/?sh=235835027bc7

Focus on Transparency/Explainability

- **Trustable models**
  - **Intelligibility:** Users need to be able to understand the link between inputs and outputs. This link, therefore, needs to be visible, rather than hidden in a black box.
  - **Predictability/Consistency:** When your user has seen results or has explored the model definition, they should be able to be in the ballpark of knowing which of two cases is more or less likely to be a particular case, or will be greater if a continuous model
  - **Reflect knowledge of the real world**: Users will frequently have significant experience of the way what is being modeled behaves in real life.

Overview Conclusion

https://www.youtube.com/watch?v=Y4b-qE9RtOk

AI Ethics Challenges : Focus on AI Explainability (XAI)

Outline

- Explainability (eXplainable AI) : Introduction
- Overview of Existing XAI Methods
  - Purposes & Forms
  - Existing techniques
    - Intrinsically Explainable Methods
    - Post-Hoc Explainability Methods
  - Implementations
  - Limits
- XAI Metrics
- Conclusion

Explainability (XAI) : Introduction

Many ML Techniques

Supervised Learning	Learn to predict a situation from provided examples	• **Regression** • **Classification**
Unsupervised Learning	Discover the underlying structures characterizing data	• **Clustering** • **Dimension reduction**
Semi Supervised Learning	Adapt a supervised model to the structure of the data provided	• **Regression** • **Classification**
reinforcement learning	Learn the actions to take based on the experience acquired	Problems that may be **« gamified »** (1)

1 : "Gamification" involves translating the problem into a situation where "an agent" is either rewarded or punished based on the action taken. The objective will therefore be to maximize the rewards at the end of the experience.

For various Use Cases / Industries

Main Use Cases

- Life time value Prediction
- Sales Prediction
- Churn Detection
- Stock Optimization
- Customer Scoring
- Purshasing Optimization
- Price Prediction
- Anomalies Detection
- Fraud Detection
- Predictive Maintenance

Industries

Finance/Banks/Insurance

Industry

Media & entertainment

Healthcare

Distribution & luxury

Explainability : Why ?

Regulatory reasons

In some sectors, regulations aim to limit the use of "black box" models,
Ex : Fair credit reporting act section 609(f)[1]

User experience / relationship

The "interpretability" of a model can greatly contribute to a better acceptability of predictions

Purpose of the application

« By design » a machine learning product may have "interpretability" as a secondary objective

(1) : https://www.ffiec.gov/exam/InfoBase/documents/02-con-fair_credit_reporting_act-000799.pdf

Explainability : Which Benefits ?

Essential for trustworthy AI

EXPLAINABILITY(XAI)

TRANSPARENCY	Ensure full information and understanding of data subjects regarding AI impacts & interactions with a machine.
EXPLICABILITY / AUDITABILITY	Provide the organization with the technical capabilities to describe, inspect and reproduce the mechanisms through which AI systems make decisions (e.g. auditability).
ACCOUNTABILITY	Ensure human responsibility regarding AI-driven decisions and notably affecting data subjects and implement the right level of governance (incl. controls) to ensure the right purpose of AI systems.
ROBUSTNESS	Ensure that algorithms are secure (cybersecurity) and robust enough to deal with errors or inconsistencies.

Explainability : a new ML Metric ?

ML Basic Metrics

Model Performance

How accurate is the model for the defined problem?

Scalability

Can the model process a huge volume of data?

Model Runtime

Does building the model take a long time?

Robustness

Are the results stable over a period of time?

New ML Metrics ?

Explainability

Can the results be interpreted in a way that humans can understand?

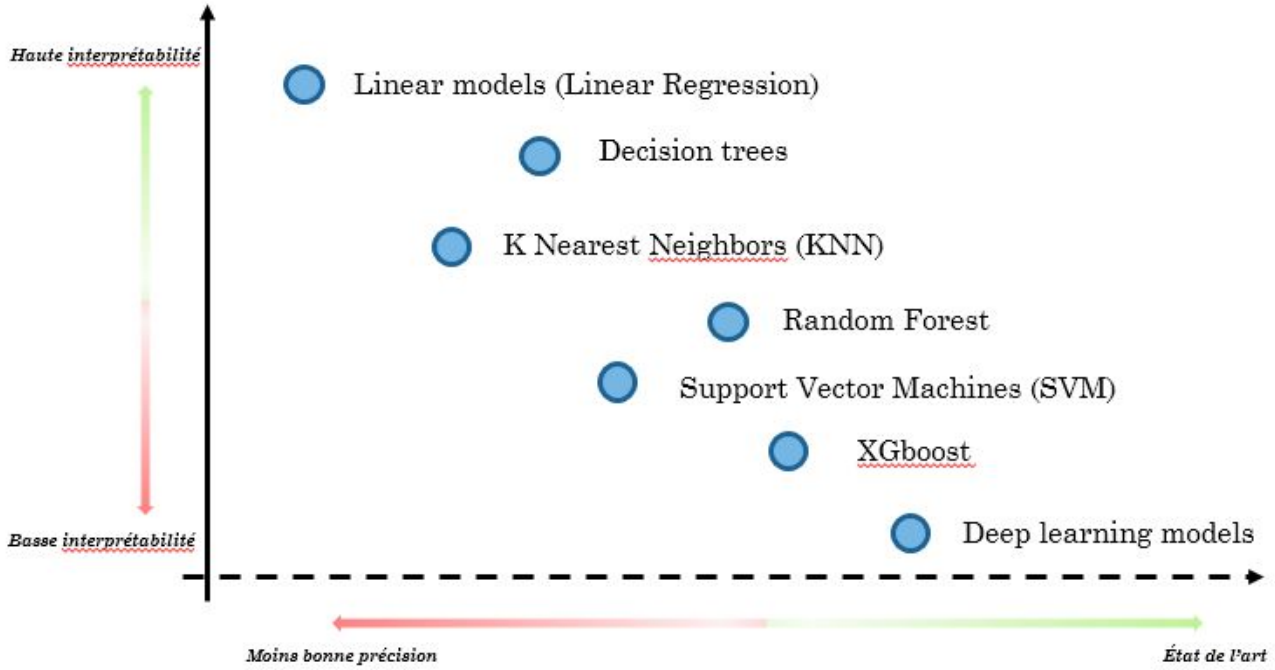
Simplicity

Can we explain the model easily?

Explainability : The trade-off

Trade-off Accuracy vs Interpretability

- The more **accurate** is a model (or a model of models, like ensemble's bagging and boosting), the more **complex** it is and so the more difficult it is to interpret it's outputs.



Explainability : for Who ?

Different Stakeholders with different needs

DATA SCIENTISTS / DATA ENGINEERS

- Improving feature engineering
- Model selection
- Improving models' robustness by checking its logic

BUSINESS EXPERTS

- Trust in the model
- Identification of new business drivers

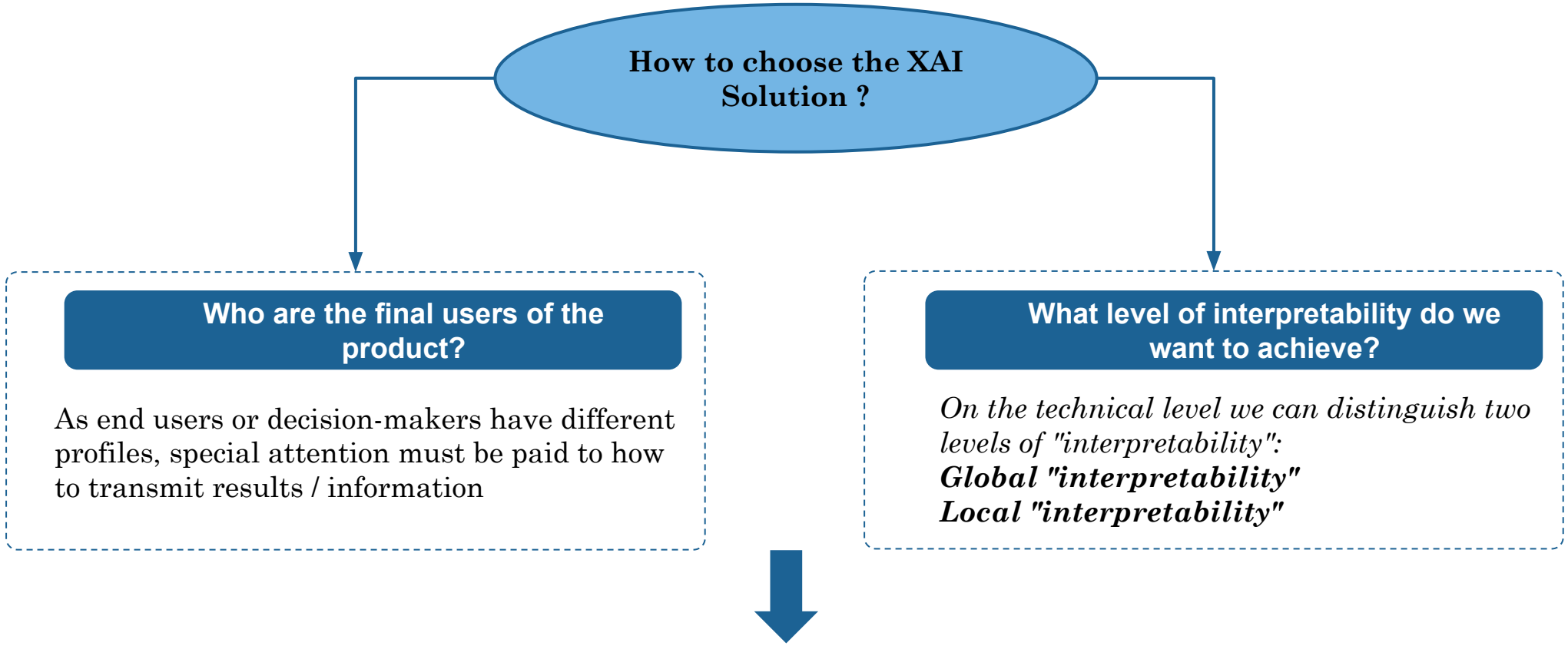
USERS (CLIENTS, EMPLOYEES, …)

- Understand decisions
- Trust in the institution using the model

REGULATORS

- Validate and authorize models
- Explain and trace complete decision making processes (Audit)

Explainability : depends on the needs



These two factors often lead to strong choices in terms of XAI development methodology.

Overview of Existing XAI Methods

-

XAI Methods : Classification

Various Criteria

Intrinsic vs Post-Hoc XAI	• Restricting the complexity of the machine learning model (intrinsic) or by applying methods that analyze the model after training (post-hoc)
XAI Outputs	• The various interpretation methods can be roughly differentiated according to their results (features statistic, features visualization, Model Internals, ..)
Model-Specific vs Model-Agnostic XAI	• Does the interpretation method specific to the model used or can be used to any machine learning model ?
Global vs Local XAI	• Does the interpretation method explain an individual prediction or the entire model behavior?

XAI Methods : Intrinsic vs Post-Hoc

- **Intrinsic Explainability :**
  - Refers to machine learning models that are considered interpretable due to their simple structure, such as decision trees or Linear models.
  - Example : **Parametrics models** (like linear regression) offer a first level of interpretation by the way of their coefficients (in some cases, this is not completely trivial !).

- **Post-hoc Explainability :**
  - Refers to the application of interpretation methods after model training. Post-hoc methods can also be applied to intrinsically interpretable models. For example, for decision trees.
  - Example **: Non-parametrics models** like tree-based models as XGBoost are more difficult to interpret because their total number of parameters is not fixed and will grow with the volume of data used for the training. Fortunately, some interpretation methods can be used (like feature importance) for helping us to understand the inner evaluation of the model for making his predictions.

XAI Methods : Ouputs

Which Insights can be expected from XAI Methods ?

- **Features summary statistic**: Many interpretation methods provide summary statistics for each feature.
  - Some methods return a single number per feature (such as feature importance),
  - Or a more complex result, which consist of a number for each feature pair (such as the pairwise feature interaction strengths) .
- **Features summary visualization**:
  - Most of the feature summary statistics can also be visualized.
  - Some feature summary are actually only meaningful if they are visualized and a table would be a wrong choice.
- **Model internals (e.g. learned weights)**:
  - Examples are the weights in linear models or the learned tree structure (the features and thresholds used for the splits) of decision trees.
  - Another method that outputs model internals is the visualization of feature detectors learned in convolutional neural networks.
- **Data point**: This category includes all methods that return data points (already existent or newly created) to make a model interpretable. Example :
  - Find a similar data point by changing some of the features for which the predicted outcome changes in a relevant way (e.g. a flip in the predicted class).
- **Intrinsically interpretable model**: One solution for interpreting black box models is to approximate them (either globally or locally) with an interpretable model.

XAI Methods : Model-Specific vs Agnostic

- **Model-specific interpretation tools :**
  - Are limited to specific model classes.
  - The interpretation of regression weights in a linear model is a model-specific interpretation, since -- by definition -- the interpretation of intrinsically interpretable models is always model-specific.
  - Tools that only work for the interpretation of e.g. neural networks are model-specific.

- **Model-agnostic interpretation tools :**
  - can be used on any machine learning model and are applied after the model has been trained (post-hoc).
  - These agnostic methods usually work by analyzing feature input and output pairs.
  - By definition, these methods cannot have access to model internals such as weights or structural information.

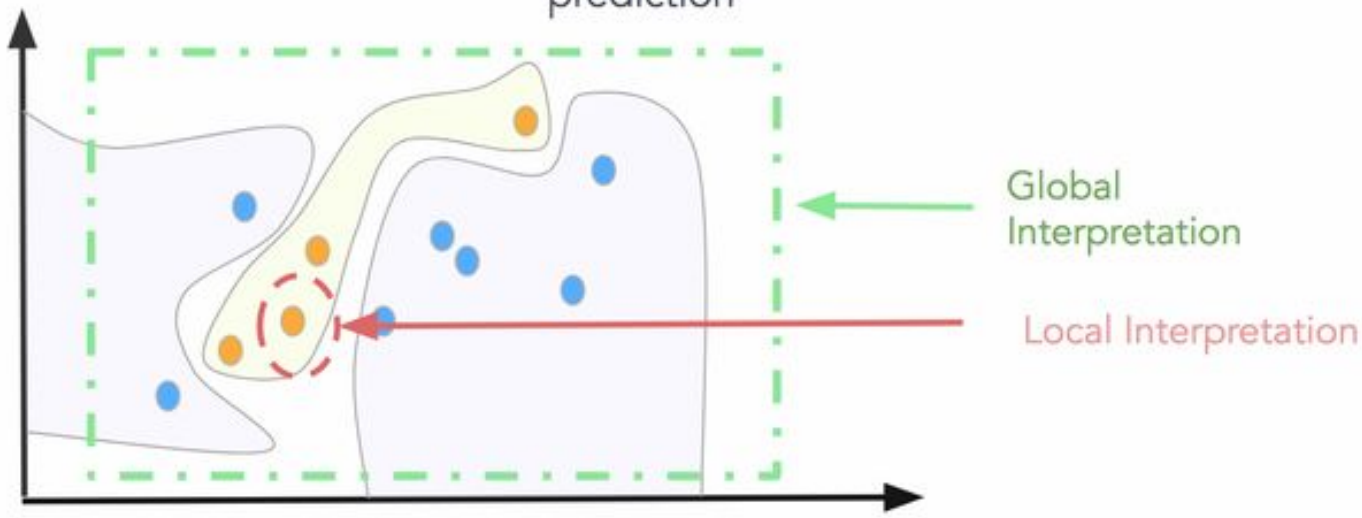
XAI Methods : Global Vs Local

Global Interpretation

Being able to explain the conditional interaction between dependent(*response*) variables and independent(*predictor, or explanatory*) variables based on the complete dataset

Local Interpretation

Being able to explain the conditional interaction between dependent(*response*) variables and independent(*predictor, or explanatory*) variables wrt to a single prediction



Source : datascience.com

XAI Methods : Global Vs Local

Explaining the model or specific decisions

Global Interpretation Methods

Model explanation : Average impact of each variable

- Understand the main decision drivers
- Detect data quality issues
- Assess engineered features effectiveness
- Verify soundness from business viewpoint

Local Interpretation Methods

Prediction explanation : Impact of each variable for individual predictions

- Understand a specific decision
- Fine understanding of model's behaviour
- Monitor model call drift over time

XAI Methods : Existing techniques

- **Intrinsically Explainable Methods**
  - White Box Models
    - Simple by Design
    - Rule based
  - Simulate BlackBox Models
- **Post-Hoc Explainability**
  - **Model Agnostic Methods**
    - Permutation Feature Importance
    - Visualisal Methods (PDP, ICE..)
    - Global Surrogate Models
    - Weight Based Methods (LIME, SHAP, ..)
    - Anchors
  - **Model Specific Methods**
    - Deep Learning Specific

XAI Methods : Intrinsically Explainable Methods

« WhiteBox Models »

If interpretability is of utmost importance for the results of the project, it is best to use an interpretable modeling technique from the start.

Low Complexity

⊕

« Simulate a Black Box Model : Advanced Feature engineering »

Depending on a "Baseline" obtained through a black box model to be defined, it is possible to improve a simple model by making it more complex iteratively.

High Complexity

⊕ ⊕ ⊕

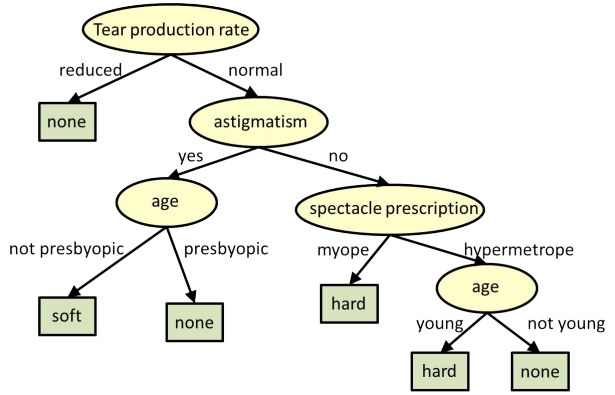
XAI Methods : Intrinsically Explainable Methods

White box

Simple by design

Many algorithms are by their construction explicable

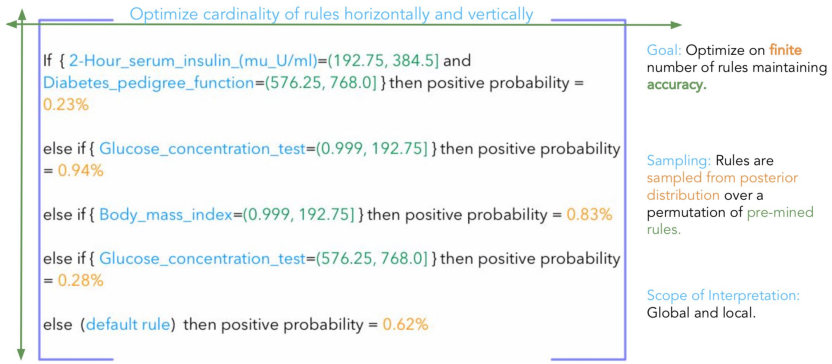
- Generalized linear models
- Decision trees
- K-Nearest Neighbors
- Etc.



« Rule based models »

Build naturally interpretable models via rule extraction:

- Scalable Bayesian Rule Lists
- RuleFit
- Arules
- Frequent Pattern Growth



XAI Methods : Intrinsically Explainable Methods

Simulate a Black Box Model

By manually synthesizing new variables :

- **Perform smart transformations on the input data**
- **Develop synthetic indicators as new training variables**
- **Include combinations of variables to simulate an interaction factor**

A Typical Workflow :

1. **Define and calculate business-specific KPIs**
2. **Create interaction variables**
3. **Consolidate the training base**
4. **Use a white box Model**

→ **A globally interpretable Model**

Strong Business involvement

XAI Methods : Post-Hoc Explainability

Permuation Feature Importance	Permutation feature importance measures the increase in the prediction error of the model after we permuted the feature's values, which breaks the relationship between the feature and the true outcome.
Visual Methods (PDP, ICE, …)	Show on a graph how the prediction of a sample would evolve if its feature values were modified. Example : “If you successively increase your monthly income by 100€, here is how the probability of getting a loan changes.”
Global Surrogate Models	Using surrogate (interpretable) models in order to explain the predictions of an algorithm around an observation
Weight Based Methods (SHAP, LIME, ..)	Assign a weight to each feature proportional to its contribution in the sample prediction. Example : “Your loan application was rejected because of your income (70% of the decision) and your young age (30%).”

XAI Methods : Post-Hoc Explainability

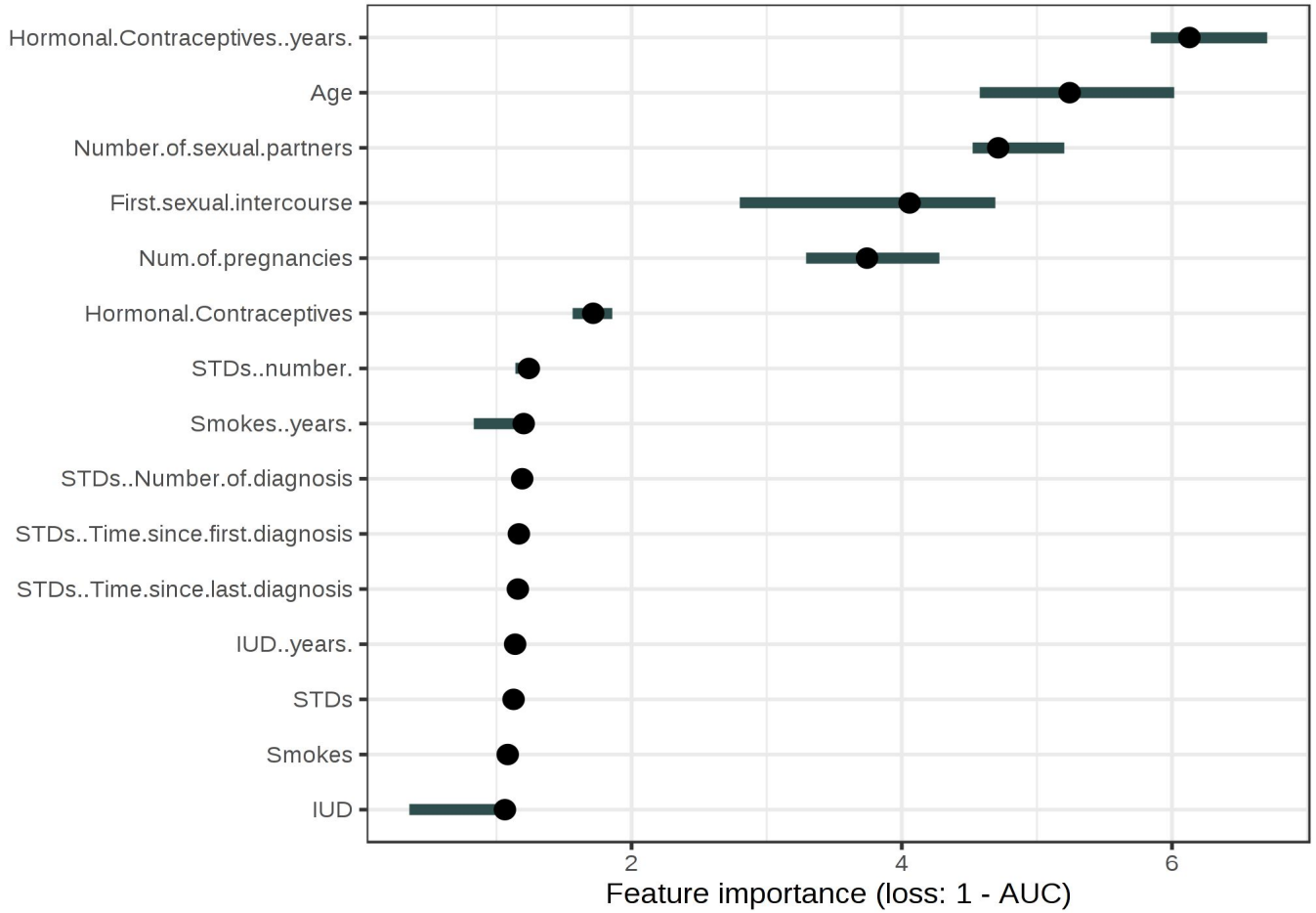
Permutation Feature Importance

- What features does a model think are important? Which features might have a greater impact on the model predictions than the others?
- This concept is called feature importance and **Permutation Importance** is a technique used widely for calculating feature importance. It helps us to see when our model produces counterintuitive results, and when it's working as we'd expect.

- Permutation Importance works for many ML Methods, the idea is simple:
  - Randomly permutate or shuffle a single column in the dataset when leaving all the other columns intact. A feature is considered "important" if the model's accuracy drops a lot and causes an increase in error. On the other hand, a feature is considered 'unimportant' if shuffling its values don't affect the model's accuracy.

XAI Methods : Post-Hoc Explainability

Permutation Feature Importance



Example : The importance of each of the features for predicting cervical cancer with a random forest.

Source : https://christophm.github.io/interpretable-ml-book/feature-importance.html

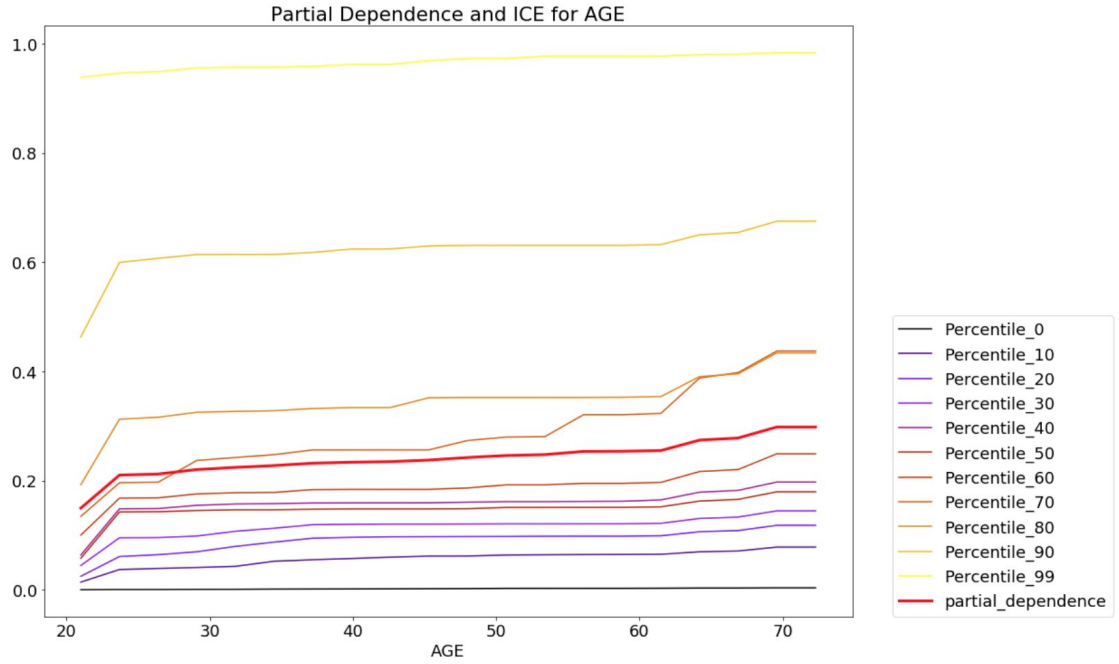
XAI Methods : Post-Hoc Explainability

Visual Methods : Partial Dependence Plots (PDP)

- The partial dependence plot (short PDP or PD plot) shows the marginal effect one or two features have on the predicted outcome of a machine learning model(J. H. Friedman 2001).

- A PDP shows whether the relationship between the target and a feature is linear, monotonous or more complex. For example, when applied to a linear regression model, partial dependence plots always show a linear relationship.

- PDP can show the relationship between the target and the selected features via 1D or 2D plots.



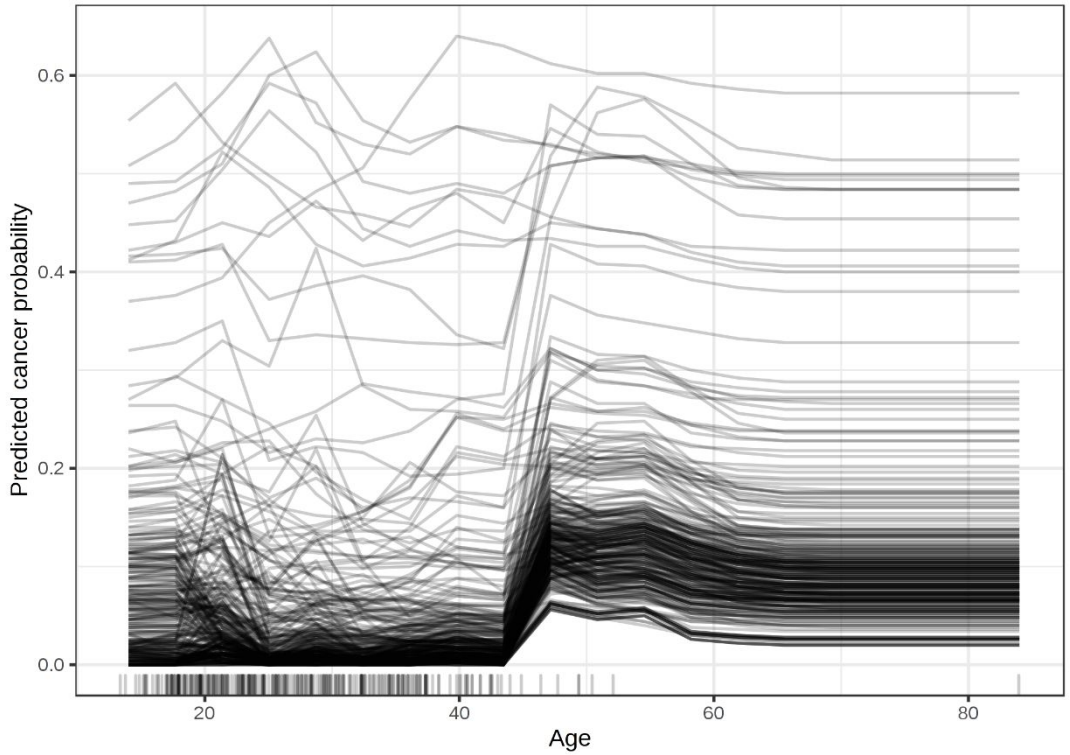
Partial dedepence plot on an AGE variable vs percentiles

While feature importance shows **WHAT** variables most affect predictions, partial dependence plots show **HOW** a feature affects predictions.

XAI Methods : Post-Hoc Explainability

Visual Methods : Individual Conditional Expectation (ICE)

- Individual Conditional Expectation (ICE) plots display one line per instance that shows how the instance's prediction changes when a feature changes.
- The PDP for the average effect of a feature does not focus on specific instances, but on an overall average. The equivalent to a PDP for individual data instances is called individual conditional expectation (ICE) plot
- An ICE plot visualizes the dependence of the prediction on a feature for *each* instance separately, resulting in one line per instance, compared to one line overall in partial dependence plots. A PDP is the average of the lines of an ICE plot.



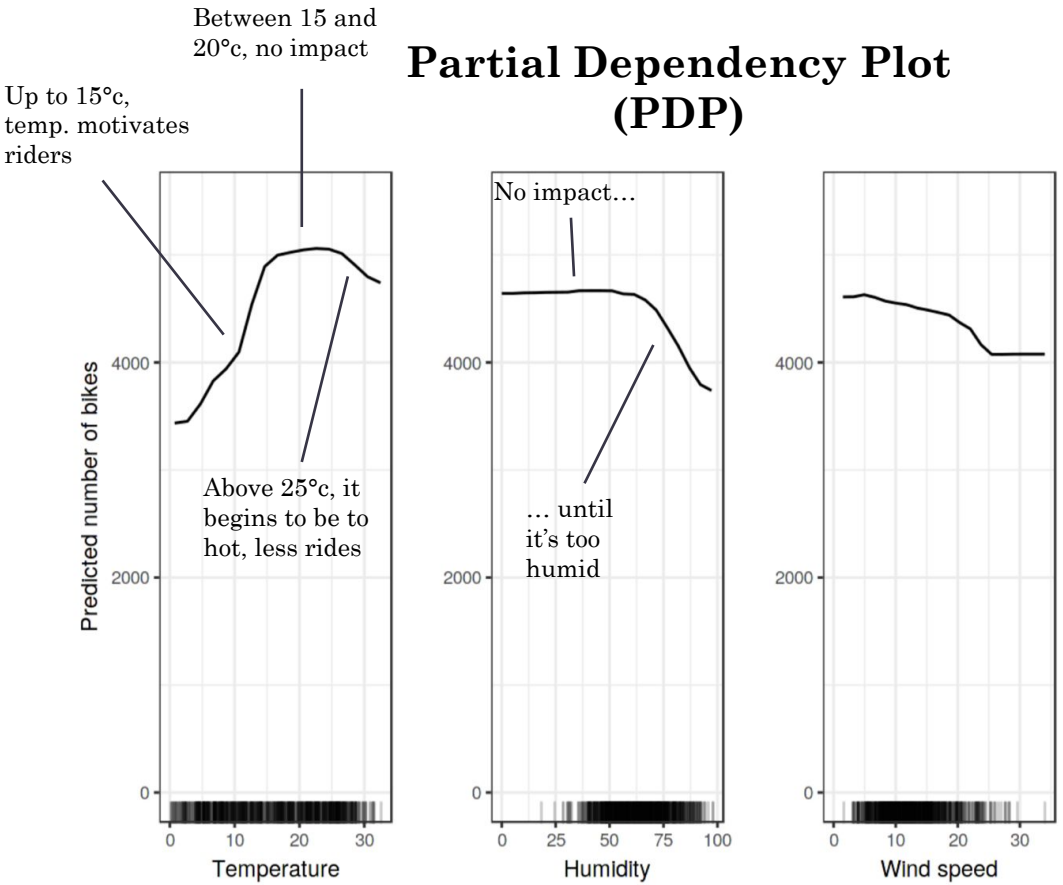
ICE plot of cervical cancer probability by age
1

(1) : https://christophm.github.io/interpretable-ml-book/ice.html

XAI Methods : Post-Hoc Explainability

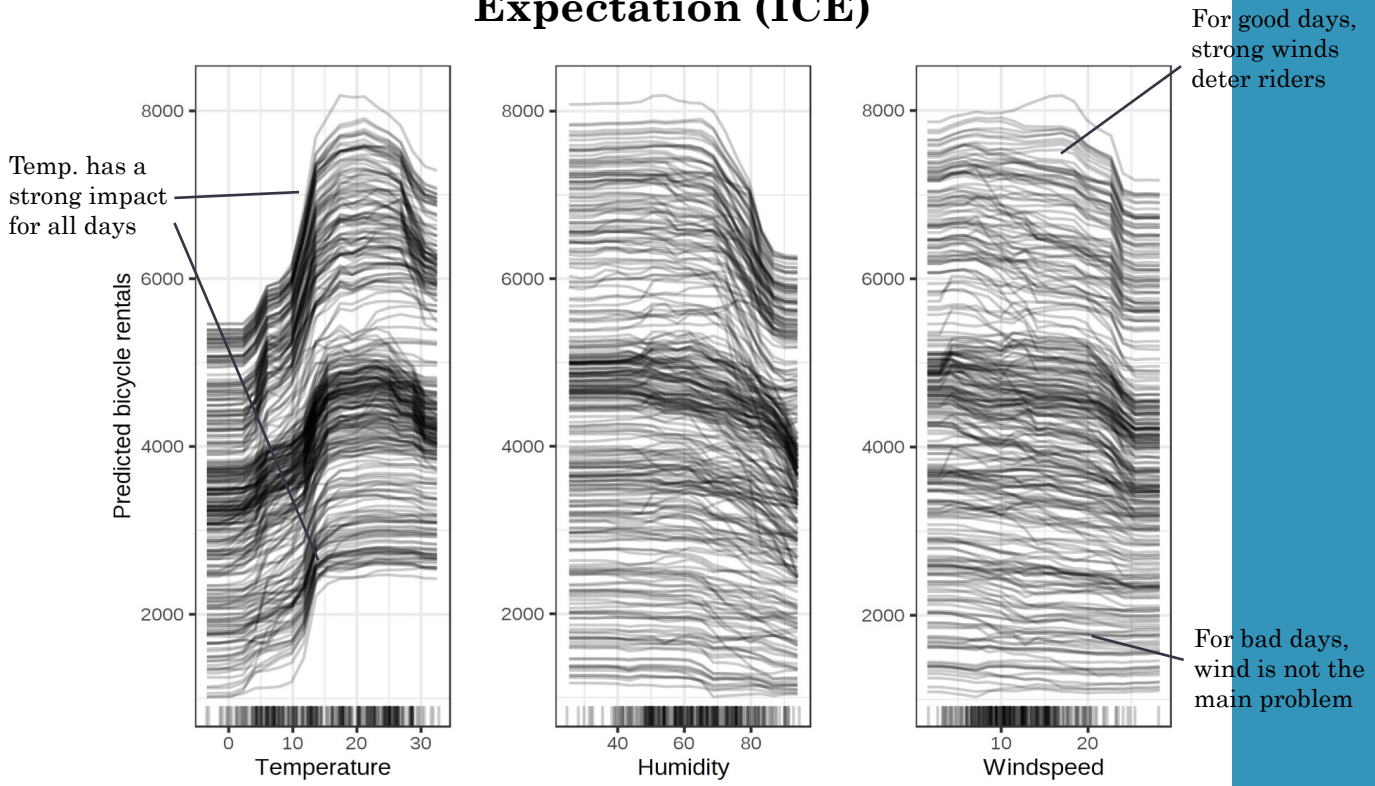
Visual Methods : PDP & ICE

Partial Dependency Plot (PDP)



1 line = **average** model outcome evolution with respect to a specific variable

Individual Conditional Expectation (ICE)



1 line = model outcome evolution, **for a given sample**, with respect to a specific variable if its value was different

Source : https://christophm.github.io/interpretable-ml-book/ice.html

XAI Methods : Post-Hoc Explainability

Global Surrogate Models

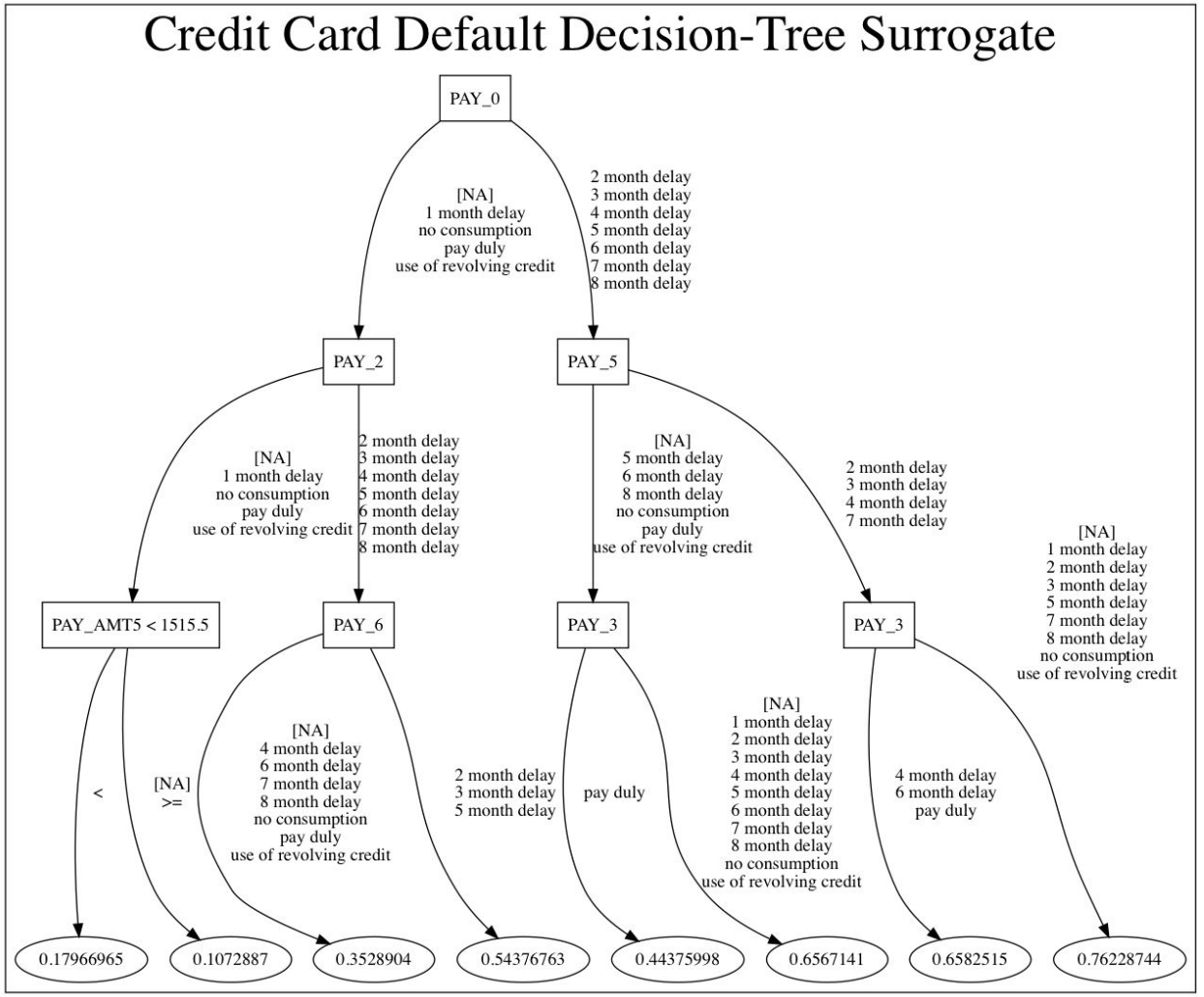
- A global surrogate model is an interpretable model that is trained to approximate the predictions of a black box model.
- We can draw conclusions about the black box model by interpreting the surrogate model (Solving machine learning interpretability by using more machine learning!)
- We want to approximate our black box prediction function "f" as closely as possible with the surrogate model prediction function "g", under the constraint that "g" is interpretable. For the function g any interpretable model can be used.
- For example a linear model: $g(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$
- Or a decision tree: $$g(x) = \sum_{m=1}^{M} c_m I\{x \in R_m\}$$

Source : https://christophm.github.io/interpretable-ml-book/global.html

XAI Methods : Post-Hoc Explainability

Global Surrogate Models

- **Example : Decision Tree Surrogate**
  - These are decision trees created on the basis of original data and predictions of an algorithm



Visualization of the optimal path to a prediction

XAI Methods : Post-Hoc Explainability

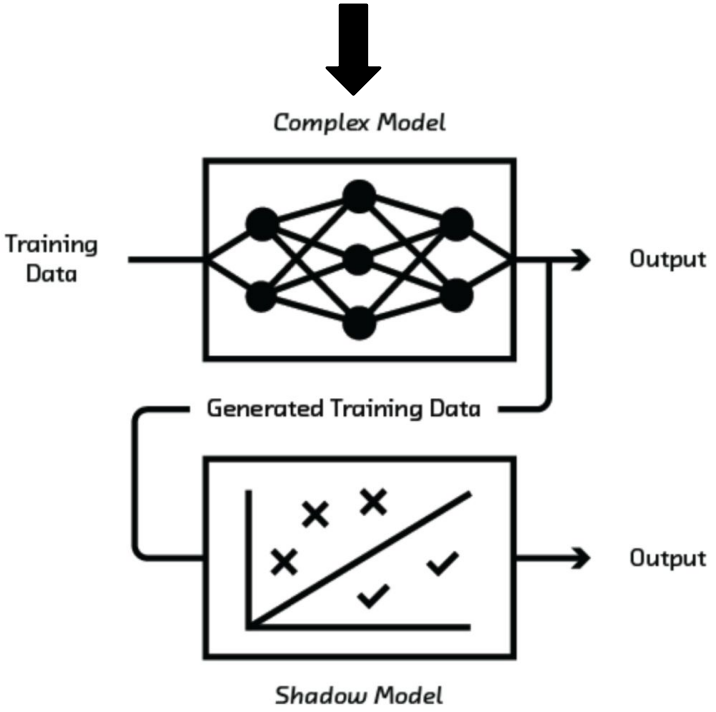
LIME

- LIME stands for:

  - **Local**: Approximate locally in the neighborhood of prediction being explained,
  - **Interpretable**: Explanations produced are human-readable,
  - **Model-Agnostic**: Works for any model like SVM, Neural networks, etc Explanations: Provides explanations of model predictions.(Local linear explanation of model behaviour)

LIME

It uses linear models to explain the decisions of an algorithm around a given observation (Locally)



Basic LIME Working

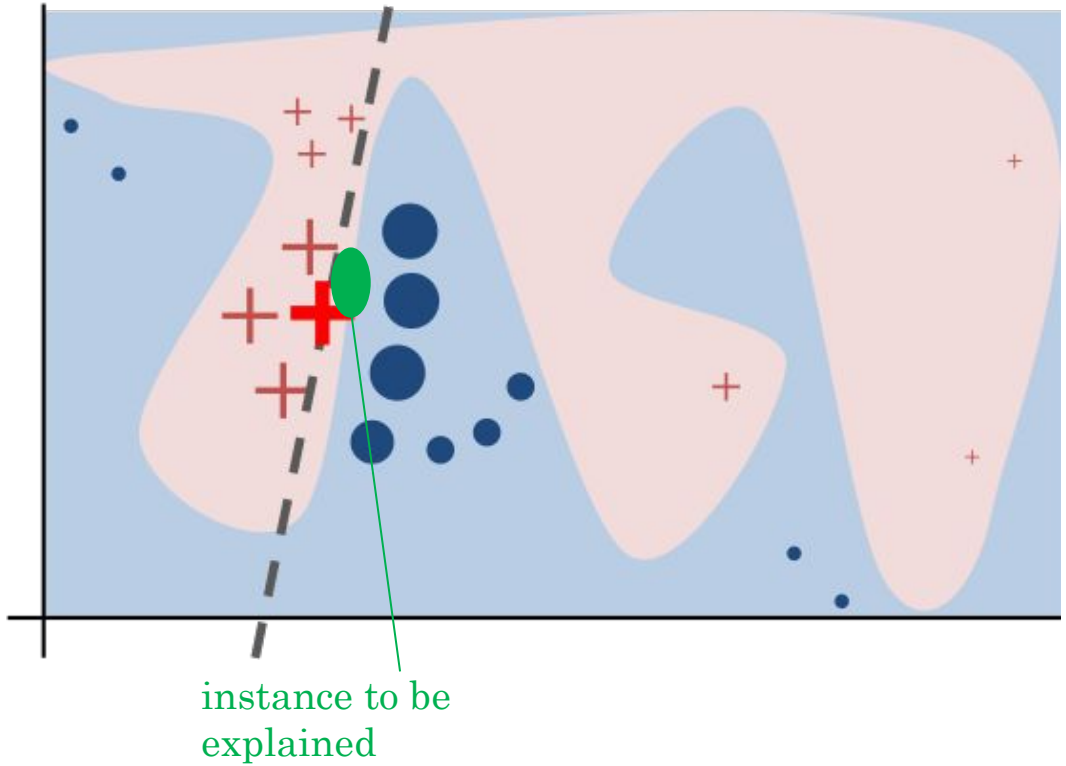
XAI Methods : Post-Hoc Explainability

LIME

- Locally approximating the model around the current prediction by a linear one and getting the variables' contribution from the linear model's coefficients (Estimate local features importance through Local linear model approximation).

LIME in 5 steps

1. Create artificial samples by adding noise on real data points
2. Get model predictions for those new points
3. Assign to each sample a weight function of its distance to the instance to explain
4. Fit a linear model (a weighted, interpretable model on the dataset with the variations)
5. The regression coefficients are the weights associated to each feature



The black box models decision function is represented by the blue/pink background. The bold red cross is the instance being explained. The grey dashed line represent the explanation model built (source : arXiv:1602.04938).

XAI Methods : Post-Hoc Explainability

LIME

The dataset is perturbed as follows :

- For **categorical data**, sampling according to the training distribution and making a binary feature (1 if the value is the same as the instance being explained, 0 otherwise)
- For **continuous data**, either :
  - Sample from a normal distribution (μ and σ comes from the training data)
  - Or transform continuous data into categorical ones (quartile binning by default)

LIME approximates the model with a Ridge regression :

$$L = \sum_{z' \in Z} [f(z') - g(z')]^2 + \sum_{j=1}^{M} \Phi_j^2$$

Black-box prediction

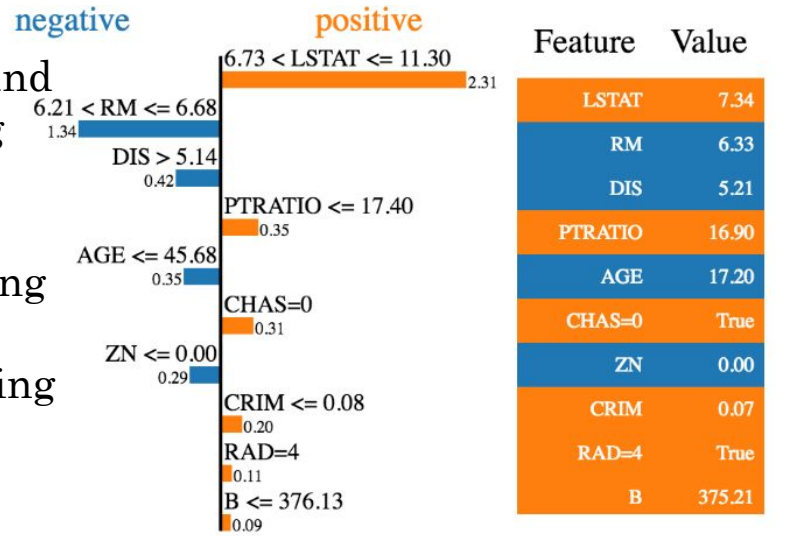
Linear function of the predictors z'_j

Regularization term

Where $g(z') = \Phi_0 + \sum_{j=1}^{M} \Phi_j * z'_j$

Coefficients of the linear model

Predictors of the perturbed dataset



Weights for each feature

Feature values of the selected instance

Note :

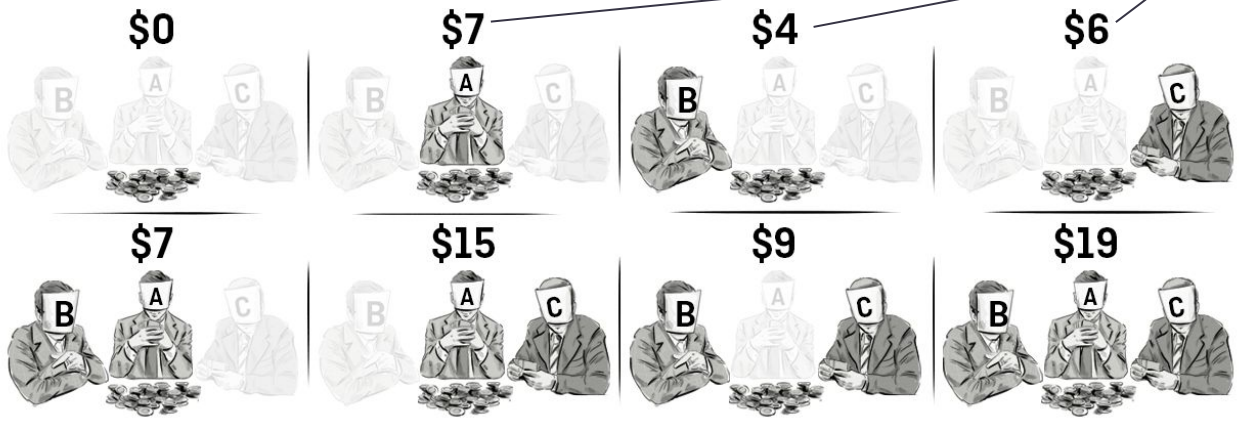
- Possibility to select the number of features in the explanation. If so, they are selected via Feature selection a priori

XAI Methods : Post-Hoc Explainability

SHAP (SHapley Additive exPlanations)

- SHAP (SHapley Additive exPlanations) is a method to explain individual predictions. SHAP is a **game theoretic** approach based on Shapley Values[1] :
  - Players = Features
  - Game = Model
  - Player's Gain/Contribution = Features importance/contribution

On the top row, A seems the most valuable



On the bottom one, it is C

How to evaluate each player's contribution in a cooperative game?

Player = Feature
Payoff = Prediction

What is the contribution of each feature to the total prediction of a ML model ?

SHAP calculates contributions such that $\Phi(A) + \Phi(B) + \Phi(C) = Total\ Prediction$

(1) Shapeley Values : https://christophm.github.io/interpretable-ml-book/shapley.html

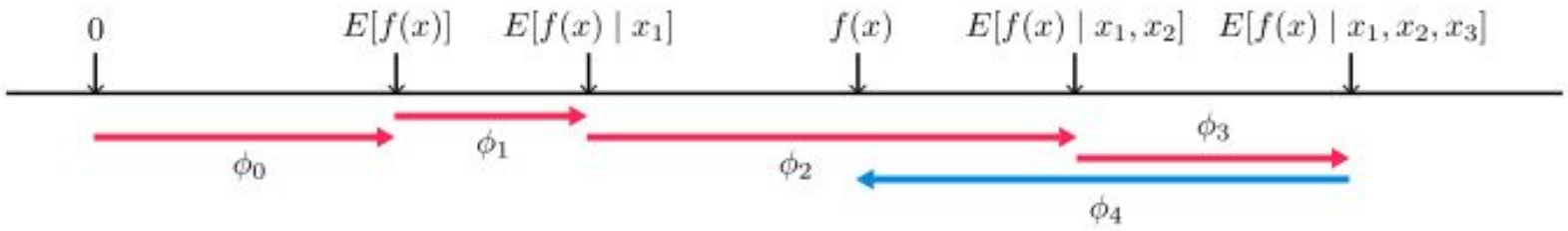
XAI Methods : Post-Hoc Explainability

SHAP (SHapley Additive exPlanations)

- An intuitive way to understand the Shapley values :
  - The feature values enter a room in random order. All feature values in the room participate in the game (= contribute to the prediction). The Shapley value "ϕij" is the average marginal contribution of feature value "xij" by joining whatever features already entered the room before, i.e.

$$\phi_{ij} = \sum_{\text{All.orderings}} val(\{\text{features.before.j}\} \cup x_{ij}) - val(\{\text{features.before.j}\})$$

- The following figure summarizes this[1] :



SHAP (SHapley Additive exPlanation) values explain the output of a function f as a sum of the effects ϕ_i of each feature being introduced into a conditional expectation. Importantly, for non-linear functions the order in which features are introduced matters. SHAP values result from averaging over all possible orderings. Proofs from game theory show this is the only possible consistent approach where $\sum_{i=0}^{M} \phi_i = f(x)$. In contrast, the only current individualized feature attribution method for trees satisfies the summation, but is inconsistent because it only considers a single ordering.

(1) : https://arxiv.org/pdf/1802.03888.pdf

XAI Methods : Post-Hoc Explainability

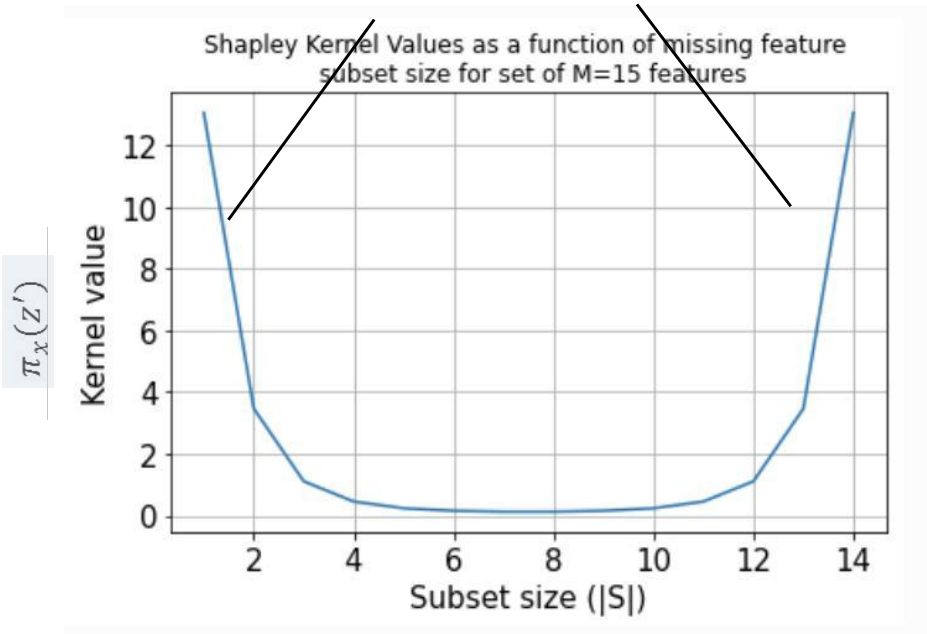
KERNEL SHAP

- Kernel SHAP is an approximation : only one model is trained and the multiple subsets are simulated through feature randomization.

- The SHAP values are calculated by solving a weighted linear regression :

    - Only some subsets are sampled
    - Each subset **z'** has a simple representation : $z'_j = 0$ if feature j is missing, 1 otherwise
    - To simulate missing features in each subset, the prediction for that subset is the average prediction among instances in a randomized dataset

More informative because we can isolate the impact of a few variables



Subset size (| z' |)

XAI Methods : Post-Hoc Explainability

TREE SHAP : An OPTIMIZED METHOD for tree-Based algorithms

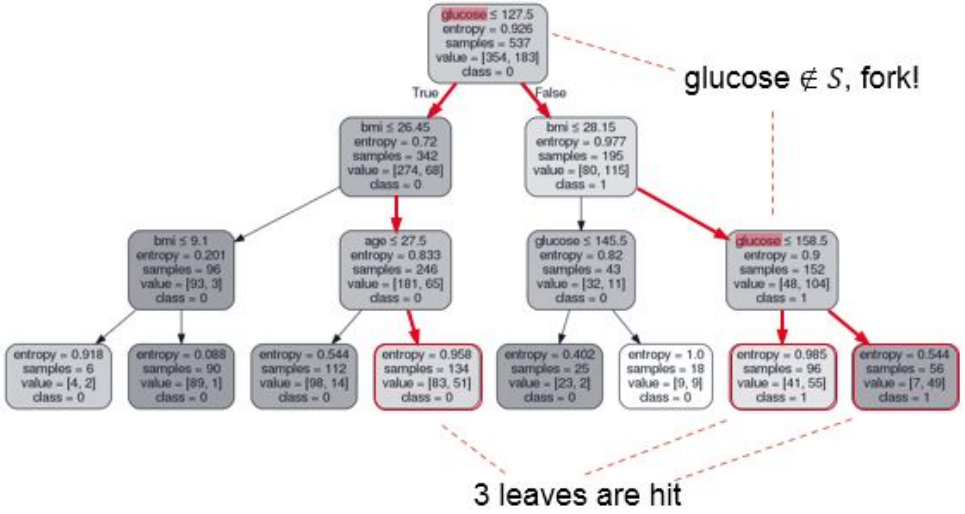
WHY TREE SHAP

- Kernel SHAP only approximates the solution as only some subsets are sampled
- Kernel SHAP is slow : simulating “missing features” takes a long time
- Tree SHAP takes advantage of the trees structure, thus we get **faster and more accurate** SHAP values

KEY IDEA : simulating missing features by fork while going down the tree

Given an instance x and a subset of present features S, the calculation of the model prediction (cf. SHAP formula) follows the decision path through the tree and :

- Return the node value if a split on a feature $i \in S$ is performed
- Take a weighted average of the values returned by children if $i \notin S$ (weighing factor = proportion of training samples having flown down each branch)



XAI Methods : Post-Hoc Explainability

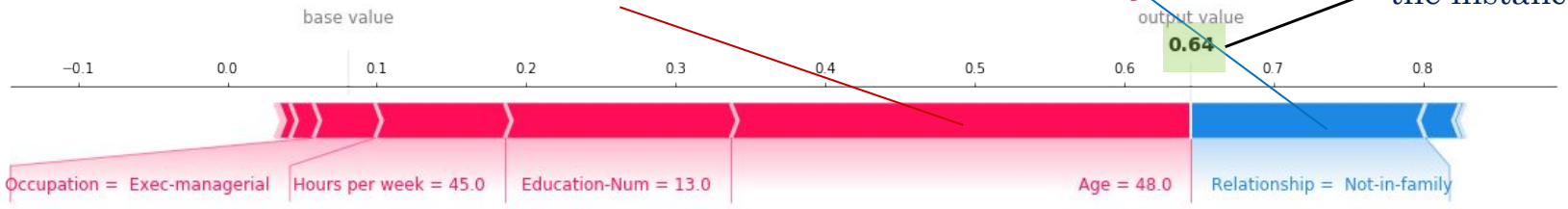
SHAP : OUTPUT

Force plot : quantify how the features drive the prediction

Relationship pushes the prediction **negatively**

Age pushes the prediction **positively**

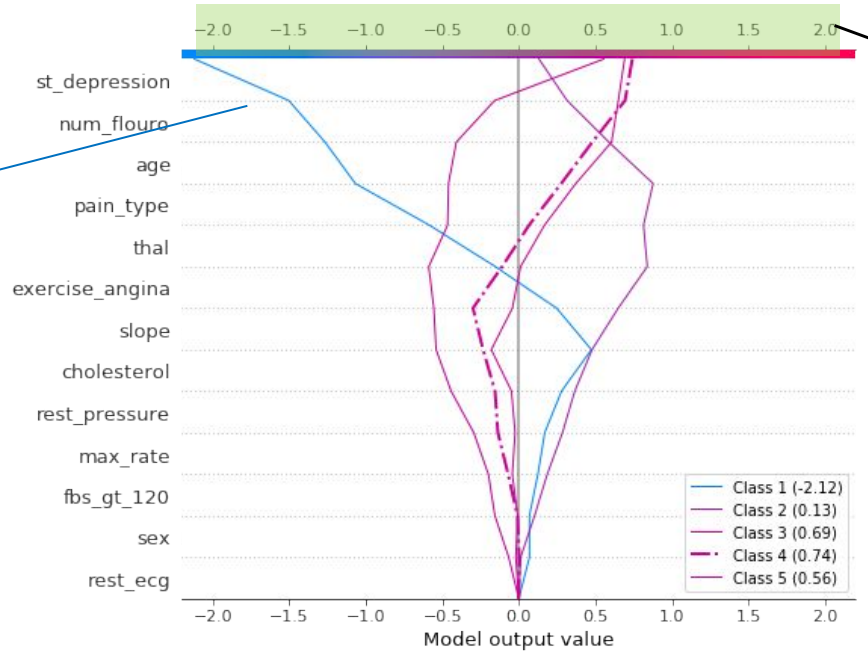
The prediction for the instance is 0.64



Decision plot : multiple instances can be displayed on a single plot

The top features have a strong negative impact on the prediction, which ends up being -2.1

The scale of the predictions



1 line = model outcome evolution for one instance

XAI Methods : Post-Hoc Explainability

ANCHORS

- Find the area including the instance to explain, in which a set of rules (anchor) leads to the same decisions as the model
- *"Other people younger than 25 and earning less that 1,000€ usually have a loan application rejected as well".*

ANCHORS

It automates the generation of clear language rules to describe a prediction of a model

	If	Predict
adult	No capital gain or loss, never married	$\leq$ 50K
	Country is US, married, work hours $>$ 45	$>$ 50K
recv	No priors, no prison violations and crime not against property	Not rearrested
	Male, black, 1 to 5 priors, not married, and crime not against property	Re-arrested
lending	FICO score $\leq$ 649	Bad Loan
	649 $\leq$ FICO score $\leq$ 699 and \$5,400 $\leq$ loan amount $\leq$ \$10,000	Good Loan

Table 3: Generated anchors for Tabular datasets

Illustration of rules produced by Achors on tests

XAI Methods : Model Specific

NEURON ACTIVATIONS VISULIZATION

- Specific to Neural Networks
- It consists of the visualization of the internal representation of the data for each layer of a Neural Network

Neuron activations visualization

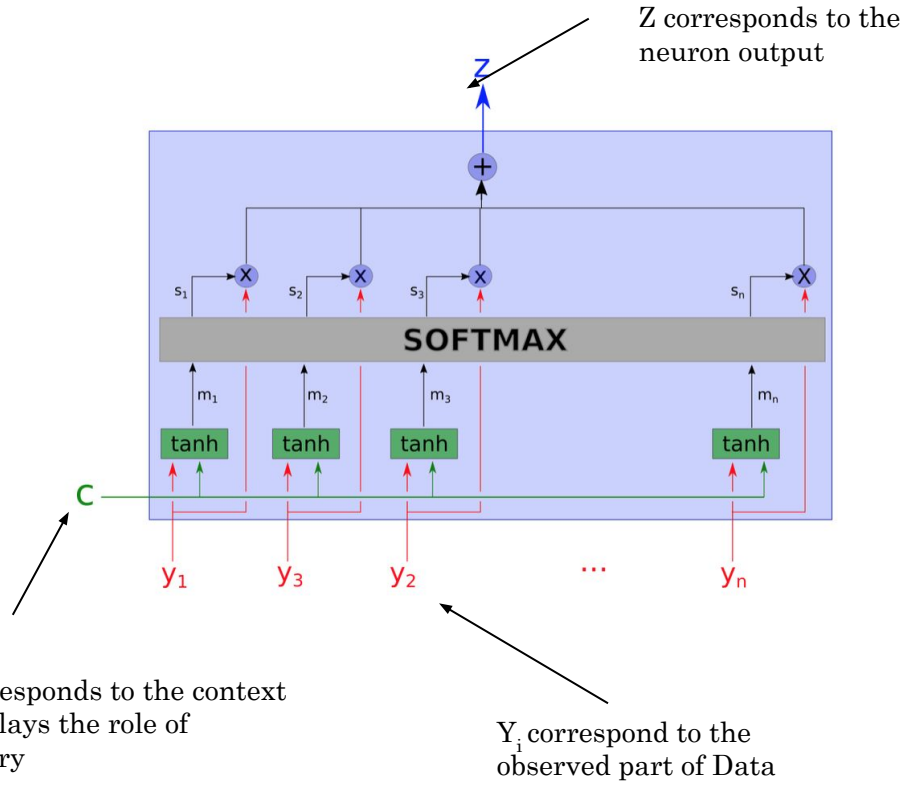
A dog seen by a Neural Network Layer

XAI Methods : Model Specific

ATTENTION MECHANISMS

- Attention mechanisms in the context of Deep Learning are techniques for directing and retaining the attention of Neural Networks on discriminating elements of the data
- Attention techniques can allocate attention, and they can learn how to do so, by adjusting the weights they assign to various inputs.
- Attention is used for machine translation, speech recognition, reasoning, image captioning, summarization, and the visual identification of objects.

Attention basic principle

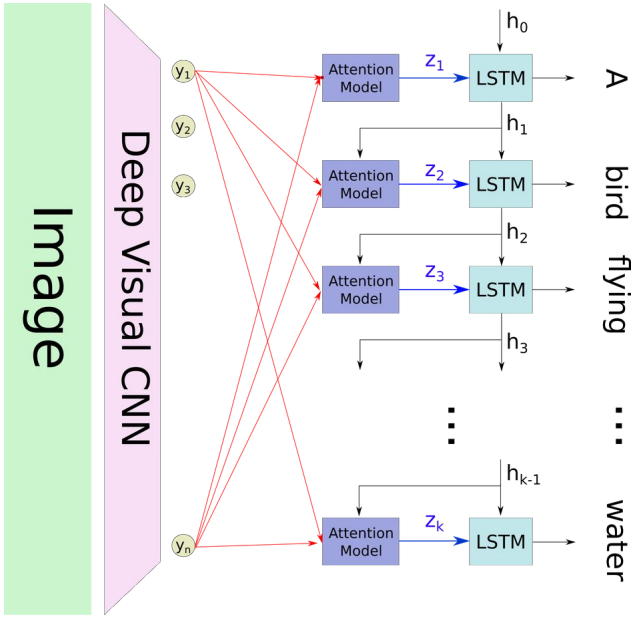


Source :
https://blog.heuritech.com/2016/01/20/attention-mechanism/

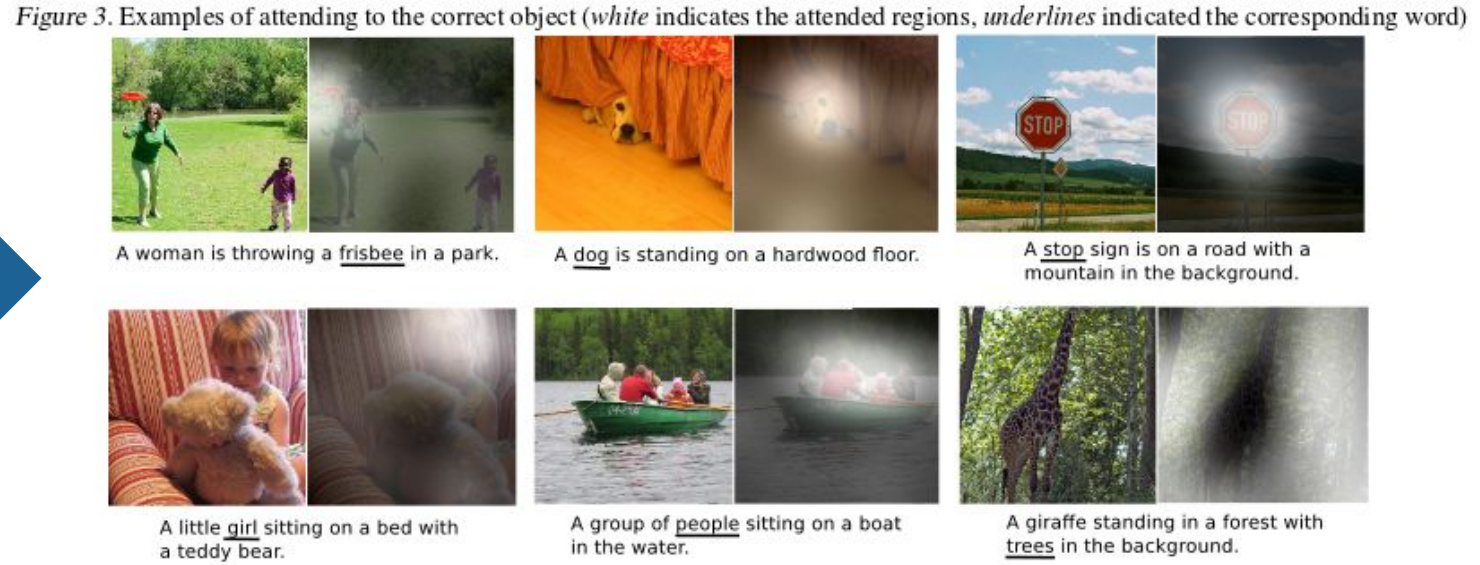
XAI Methods : Model Specific

ATTENTION MECHANISMS : Example of “Image Captioning” problem

« Image Captioning » problem architecture



Example of Attention methods on a « captioning » problem



Source :
https://blog.heuritech.com/2016/01/20/attention-mechanism/

XAI Methods : Implementation

ELI5 : for intrinsically interpretable models

- **ELI5** ("Explain Like I'm 5") is a Python library which helps to debug machine learning classifiers and regressors and explain their predictions in an easy to understand an intuitive way.
- ELI5 is a good starting point and support tree-based and parametric/linear models and also text processing and HashingVectorizer utilities from scikit-learn but **doesn't support true model-agnostic interpretations**.

LIME

LIME is implemented in Python (lime library) and R (lime package and iml package) and is easy to use.

SHAP

- SHAP values can explain the output of any machine learning model but for complex ensemble models it can be slow.
- SHAP has implementations supporting XGBoost, LightGBM, CatBoost, and scikit-learn tree models (https://github.com/slundberg/shap).

XAI Methods : Limits

LIME LACKS LOCAL ACCURACY

The linear model does not match the model output on the instance itself

$$f(x) \neq \sum_i \Phi i$$

Where :

$f(x)$ is the model output for instance x

Φi is the weight for feature i

SHAP MAY EXHIBIT UNEXPECTED BEHAVIOR ON MODELS WITHOUT A SIGNIFICANT ADDITIVE COMPONENT

For example, consider a model given by

$$f(x) = \prod_{j=1}^{d} x_j$$

where the features are independent and centered

➔ SHAP would always assign the same weights to all features

NO PERFECT WAY TO HANDLE CORRELATED VARIABLES IN SHAP

Missing features are simulated through a sampling process and SHAP offers multiple ways of doing it.

Impossible to be both “true to the data” and “true to the model”.

APPENDIX