Simple interpretability methods for black-box machine learning systems

Adriano Koshiyama

September, 2020

Table of Contents

Key Components of AI Assessment

- What do we mean by Algorithmic Impact (AI) Assessment ?
- Assessment vs by Design in AI Assessment

What do we mean by AI Assessment

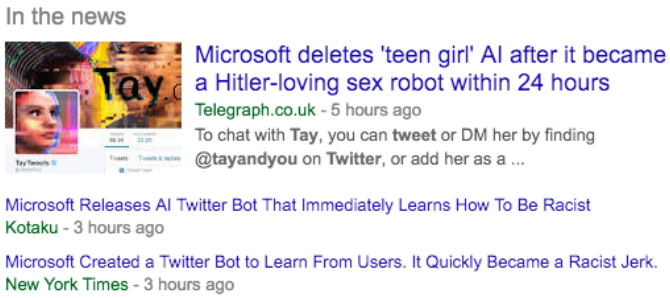
- **Algorithmic Impact Assessment** focus on **evaluating an Automated Decision-making system** mainly from a **Robustness, Fairness and Explainability** point of view
- The **goals** of AI Assessment are
  - Set the boundary, usage and shelf-life of a system
  - Build trust between the stakeholders of a system
  - Be the entry point to hold the system's creators accountable of the results of its decision-making
- We should also mention **other areas** of AI Assessment, such as **Transparency**, **Accountability**, etc.

What do we mean by AI Assessment

In a nutshell

- **Robustness**: systems should be safe and secure, not vulnerable to tampering or compromising of the data they are trained on.
- **Fairness**: systems should use training data and models that are free of bias, to avoid unfair treatment of certain groups.
- **Explainability**: systems should provide decisions or suggestions that can be understood by their users and developers.

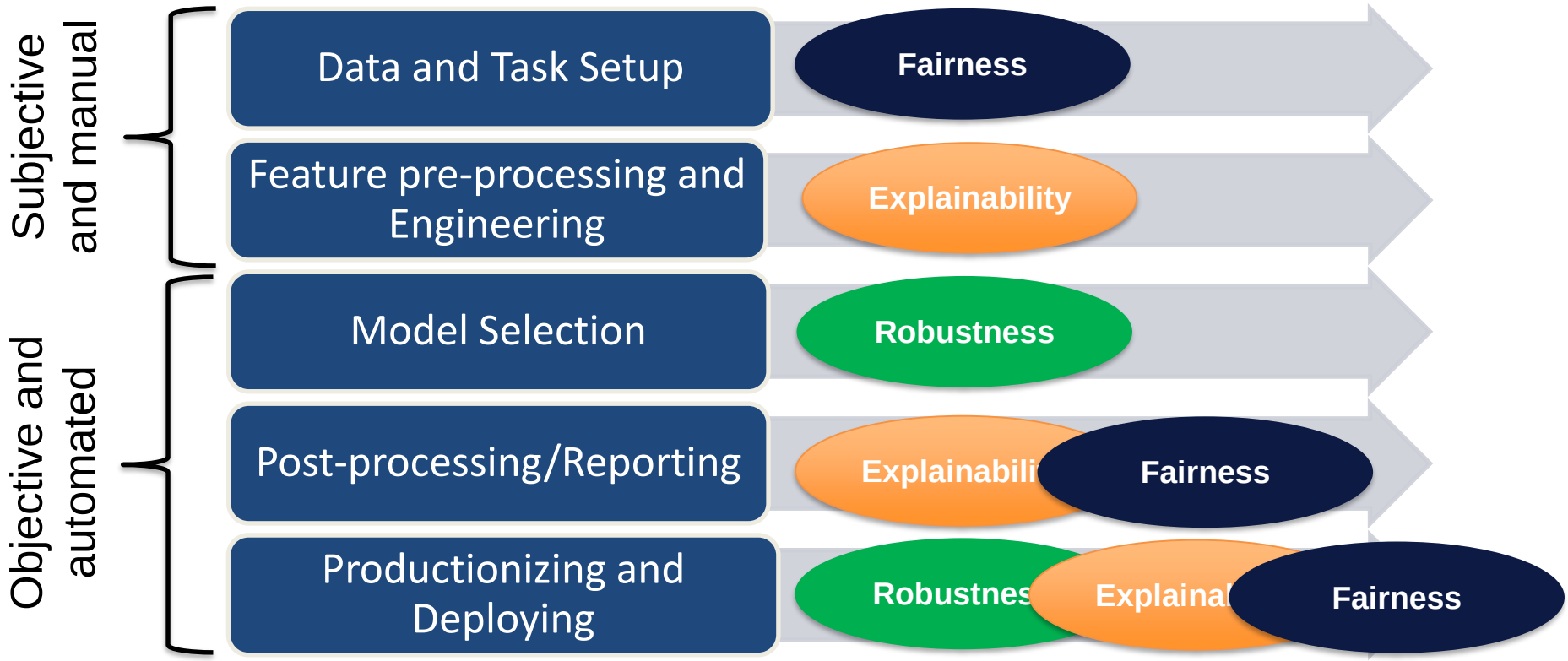
To avoid these cases



Assessment vs Modelling

From an Assessment point of view

- Areas in the modelling pipeline where a AI Assessment Analyst (AI^2) should analyse using the different criteria



Assessment vs Modelling

From a by Design point of view

- It is possible to have systems that by design are able to increase or fulfil the stakeholders demand for Fairness, Robustness and Explainability

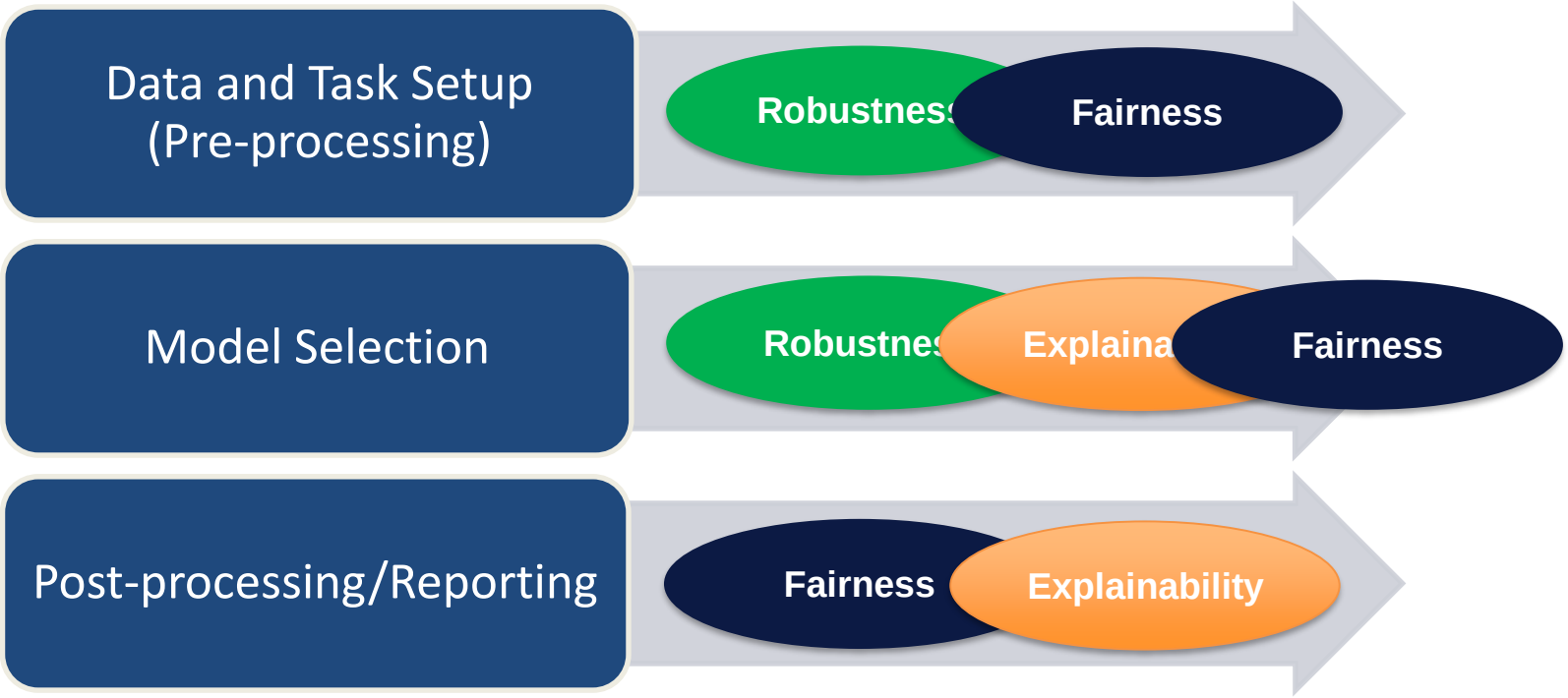


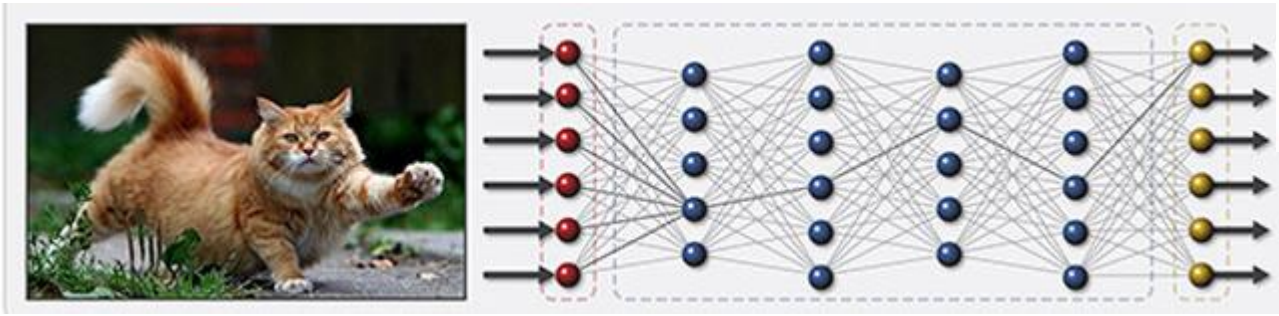
Table of Contents

Importance of Algorithmic Explainability

- What do we mean by an explainable decision
- Why and what type of Explainability
- Legal basis for Explainability
- Different types of Explainability
- Technological solutions for Explainability
- Explainability: an AI Assessment checklist
- Further reading

What do we mean by an explainable decision

Object recognition



This is a cat.

This is a cat:
- It has fur, whiskers, and claws.
- It has this feature:


Healthcare

predict breast cancer

NHS

Home | About Predict | Predict Tool | Contact | Legal | Earlier versions

What is Predict?

Predict is an online tool that helps patients and clinicians see how different treatments for early invasive breast cancer might improve survival rates after surgery.

It is endorsed by the American Joint Committee on Cancer (AJCC).

Start Predict

Did you mean to visit Predict Prostate?

Finance

Sorry, your loan application has been rejected.

If instead you had the following values, your application would have been approved:

- MSinceOldestTradeOpen: **161**
- NumSatisfactoryTrades: **36**
- NetFractionInstallBurden: **38**
- NumRevolvingTradesWBalance: **4**
- NumBank2NatlTradesWHighUtilization: **2**

(b) Counterfactual explanation

Why and what type of Explainability

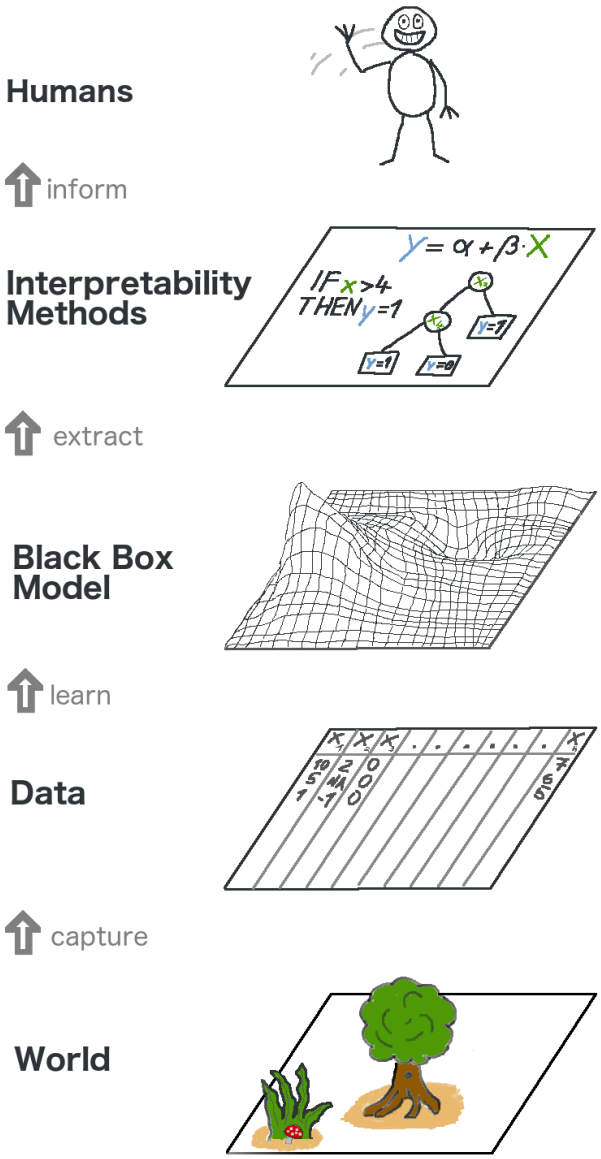
- Explicability is **crucial for building and maintaining users' and designers' trust** in AI-based decisions
  - Users: contest decisions, learning
  - Creators: knowledge discovery, debugging systems, uncover unfair decisions
- Hence, the capabilities and purpose of AI systems should be
  - **openly communicated**
  - decisions **explainable** to those **directly and indirectly affected**
  - **timely and adapted to the expertise** of the stakeholder concerned (e.g. layperson, regulator or researcher)

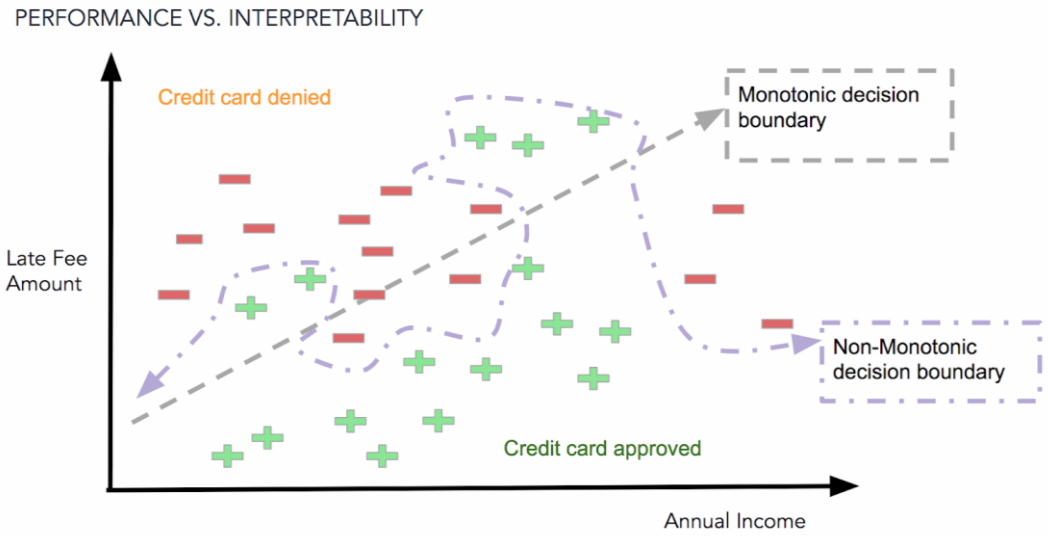
Legal basis for Explainability

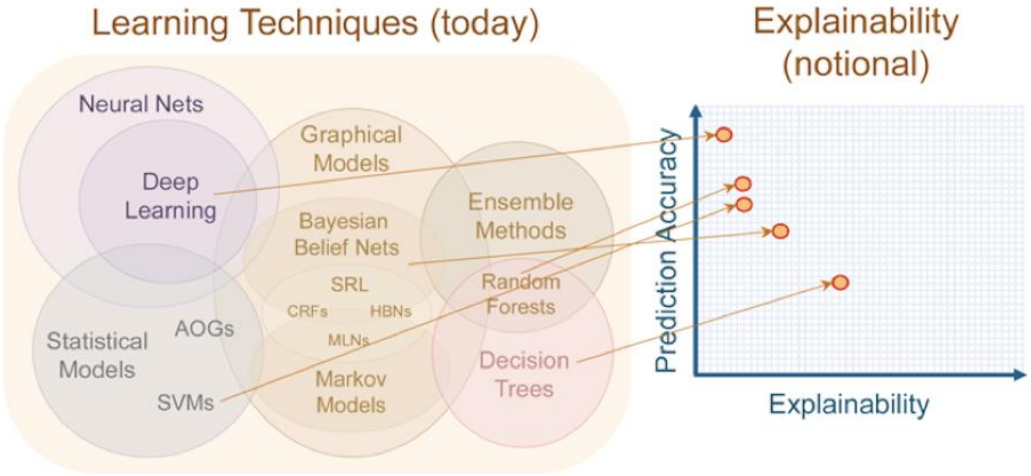
- **Credit Scoring in the US** have a well-established right to explanation
  - The Equal Credit Opportuniy Act (1974)
- Credit agencies and data analysis firms such as FICO comply with this regulation by providing a list of reasons (generally at most 4, per interpretation of regulations)

- **From an AI standpoint**, there are new regulations that gives the system's user the right (?) to know why a certain automated decision was taken in a certain form
  - **Right to an Explanation – EU General Data Protection Regulation (2018)**

Different types of Explainability

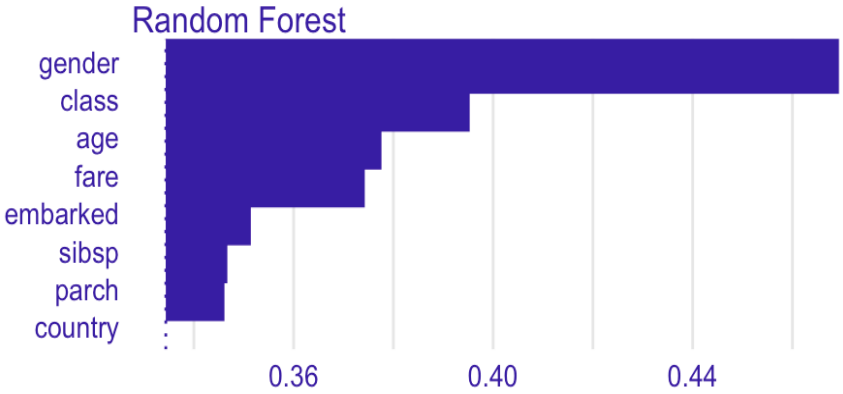






Different types of Explainability

(b) Counterfactual explanation

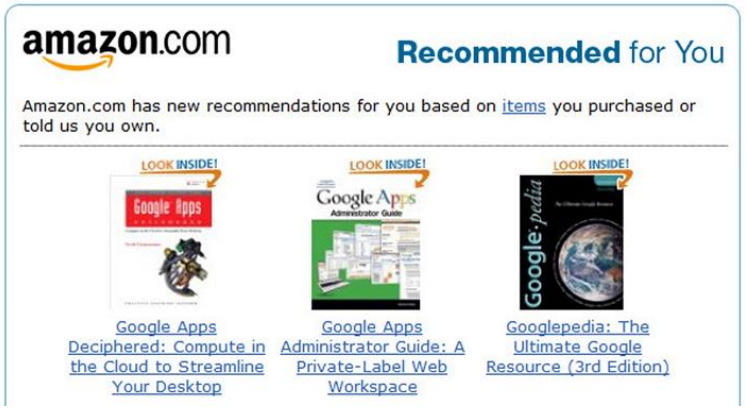


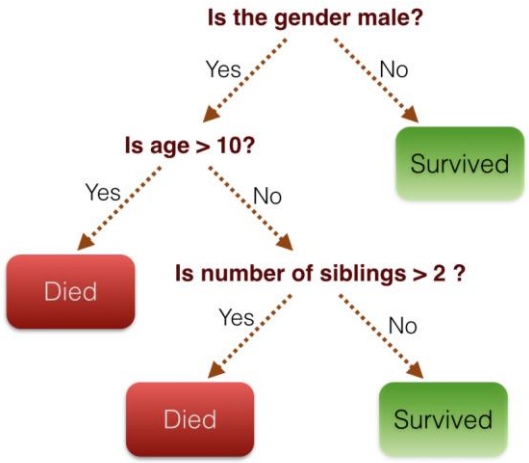
Model-agnostic

local

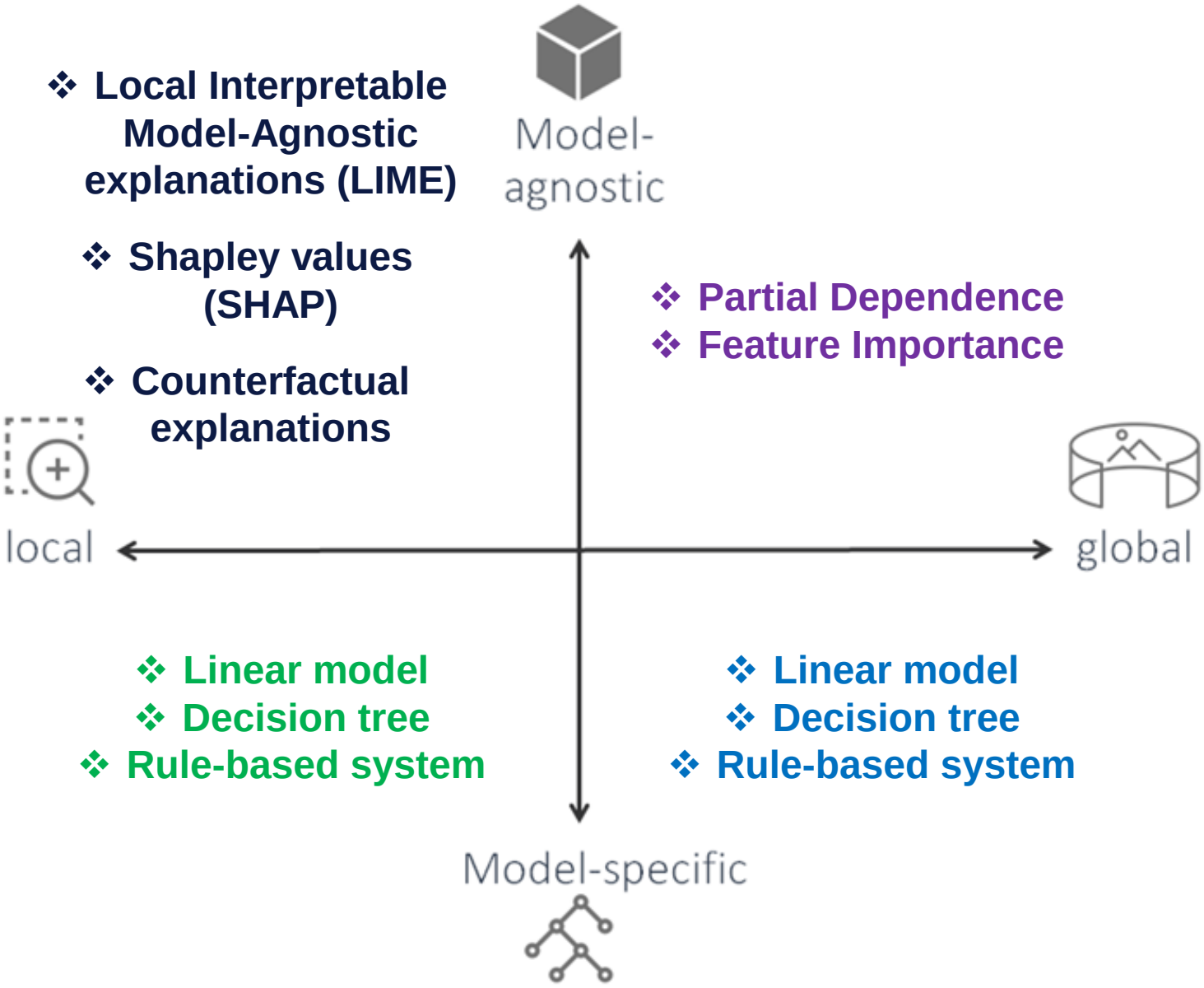
global

Model-specific



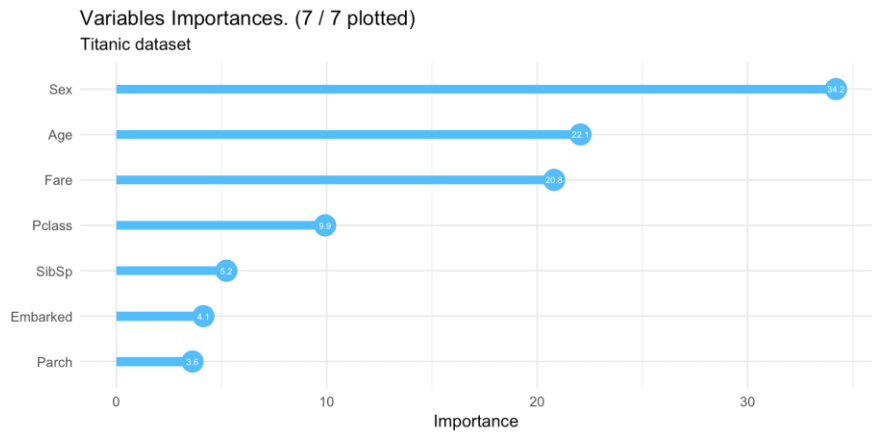


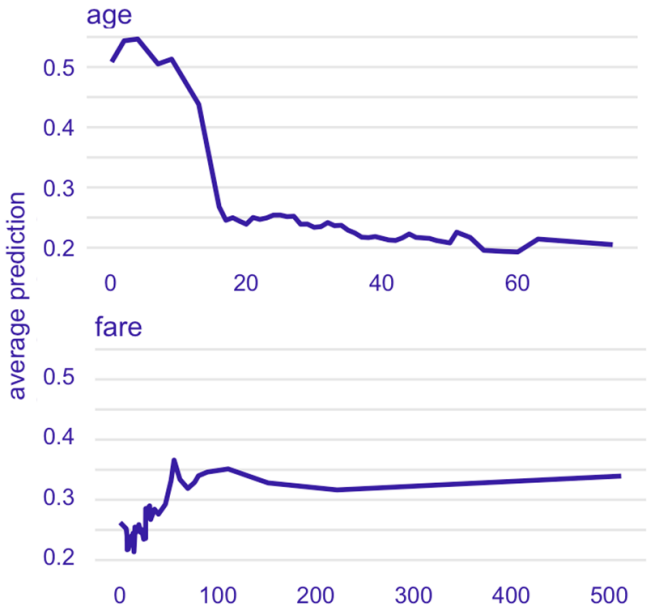
Technical solutions for Explainability



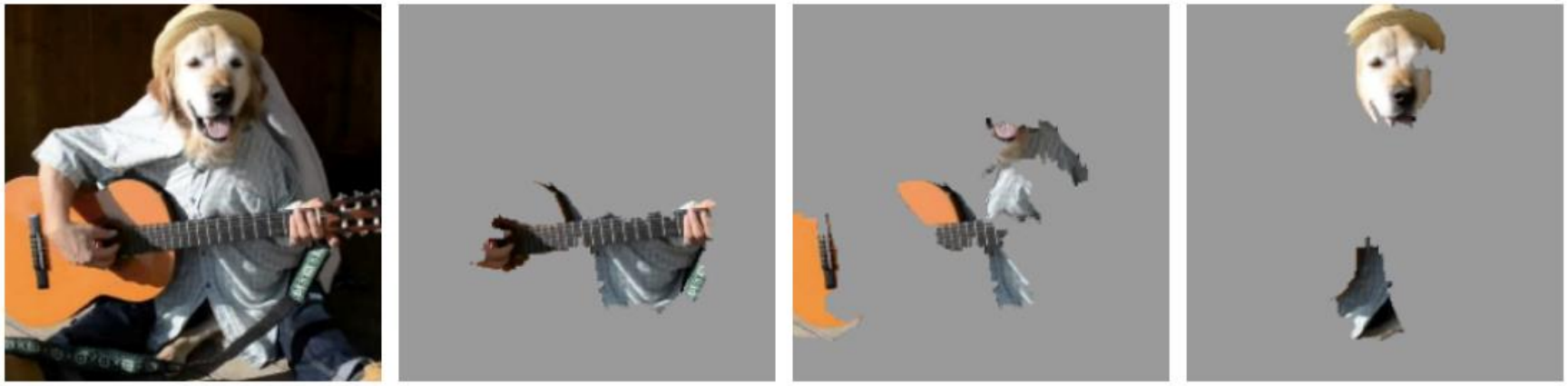
Technical solutions for Explainability

- Some examples of how it works
  - Feature and partial dependency plots





  - Local Interpretable Model-Agnostic explanations



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Explainability: an AI assessment checklist

Explainability

- Did you assess to what extent the decisions and hence the outcome made by the AI system can be understood?
- Did you ensure an explanation as to why the system took a certain choice resulting in a certain outcome that all users can understand?
- Did you design the AI system with interpretability in mind from the start?
- Did you research and try to use the simplest and most interpretable model possible for the application in question?
- Did you assess whether you can examine interpretability after the model's training and development, or whether you have access to the internal workflow of the model?

Further reading

- Legislation
  - The Equal Credit Opportunity Act: https://www.justice.gov/crt/equal-credit-opportunity-act-3
  - GDPR: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/rights-related-to-automated-decision-making-including-profiling/
- Papers and books
  - Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Model-agnostic interpretability of machine learning." arXiv preprint arXiv:1606.05386 (2016).
  - Hall, Patrick. "On the Art and Science of Machine Learning Explanations." arXiv preprint arXiv:1810.02909 (2018).
  - Hall, Patrick, and Navdeep Gill. Introduction to Machine Learning Interpretability. O'Reilly Media, Incorporated, 2018.
  - Wachter, Sandra, Brent Mittelstadt, and Chris Russell. "Counterfactual explanations without opening the black box: Automated decisions and the GDPR." Harvard Journal of Law & Technology 31, no. 2 (2017): 2018.

Further reading

- Tools
  - https://pair-code.github.io/what-if-tool/
  - https://github.com/marcotcr/lime
  - https://github.com/microsoft/interpret
  - https://github.com/slundberg/shap
- Other good online resources
  - https://christophm.github.io/interpretable-ml-book/
  - https://distill.pub/2018/building-blocks/
  - https://www.oreilly.com/ideas/ideas-on-interpreting-machine-learning

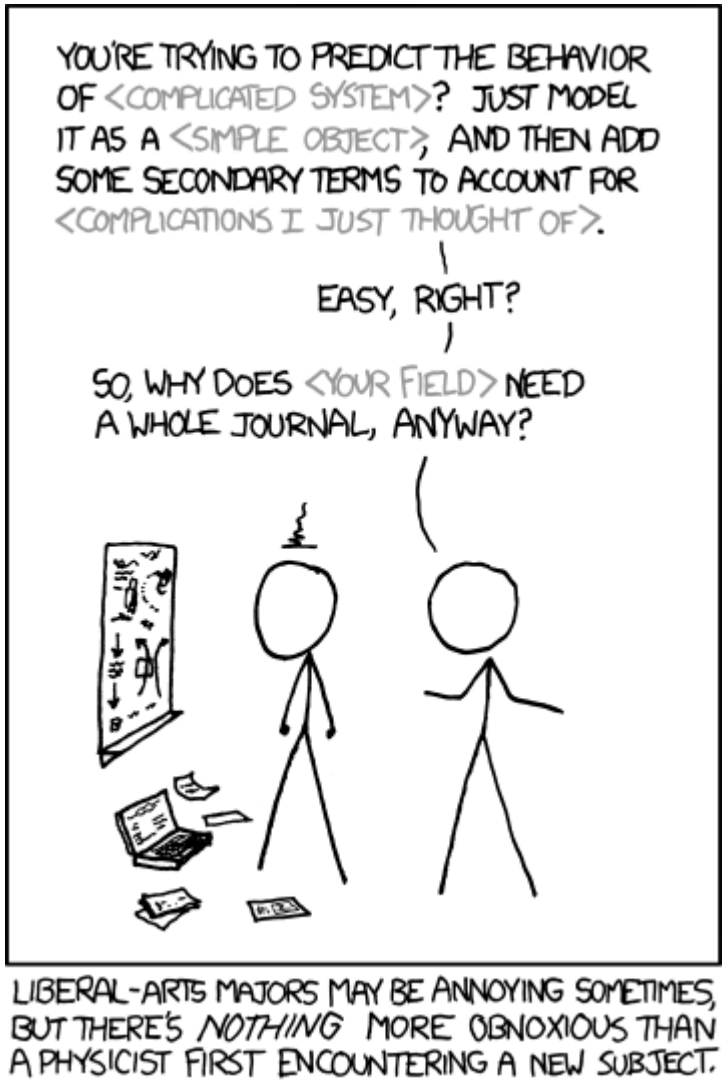


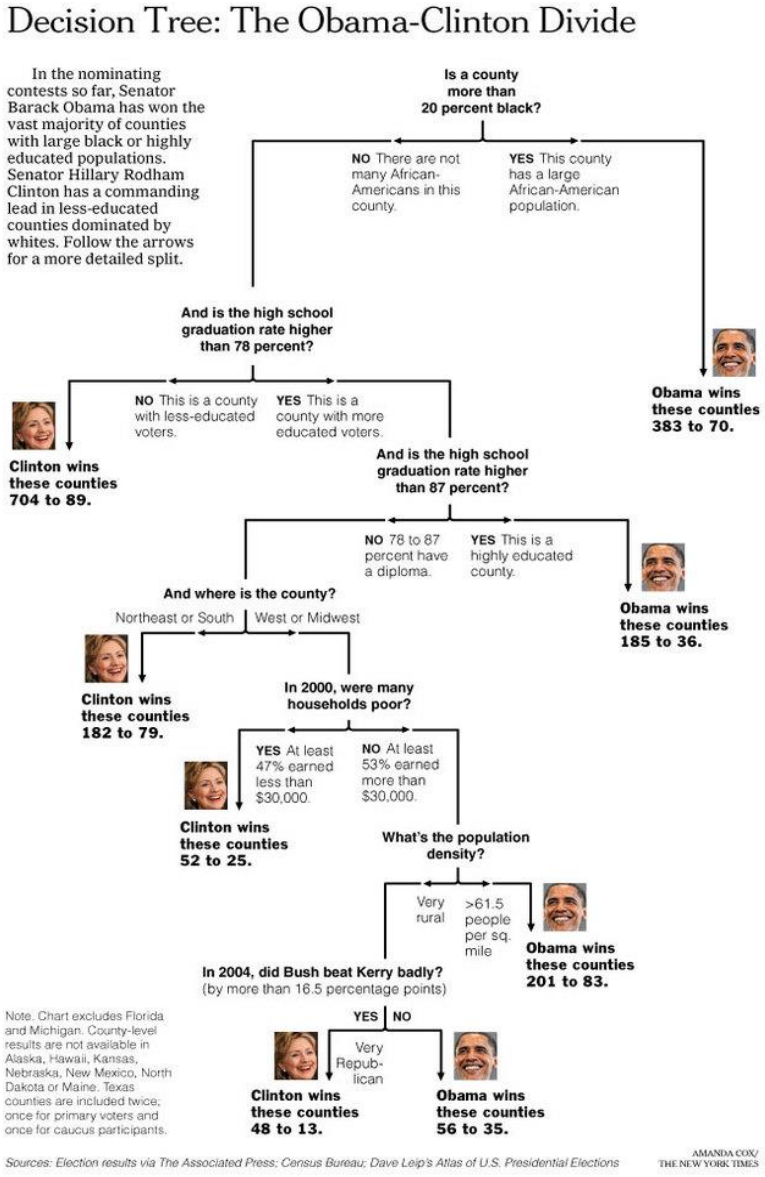
Table of Contents

Algorithmic Explainability

- Model-specific explanations: Decision Trees
  - Global explanations: tree structure
  - Local explanations: decision rule
- Model-agnostic explanations: Neural Networks
  - Global explanations: feature importance and partial dependence plots
  - Local explanations: Local Interpretable Model-agnostic Explanations (LIME)

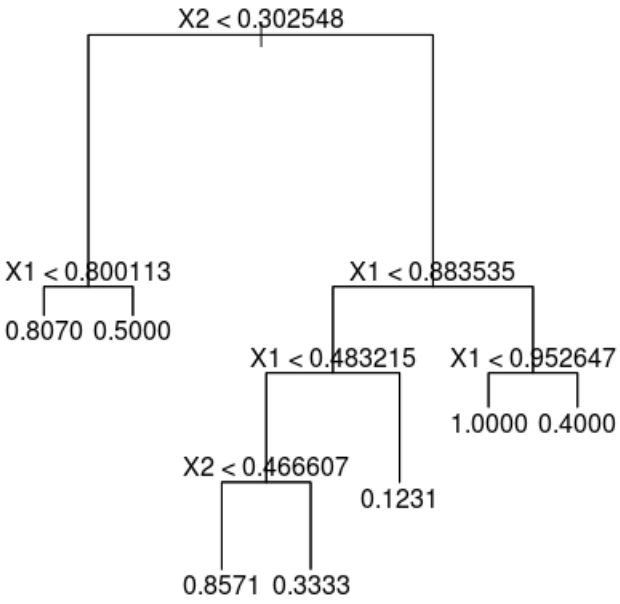
Model-specific: Global

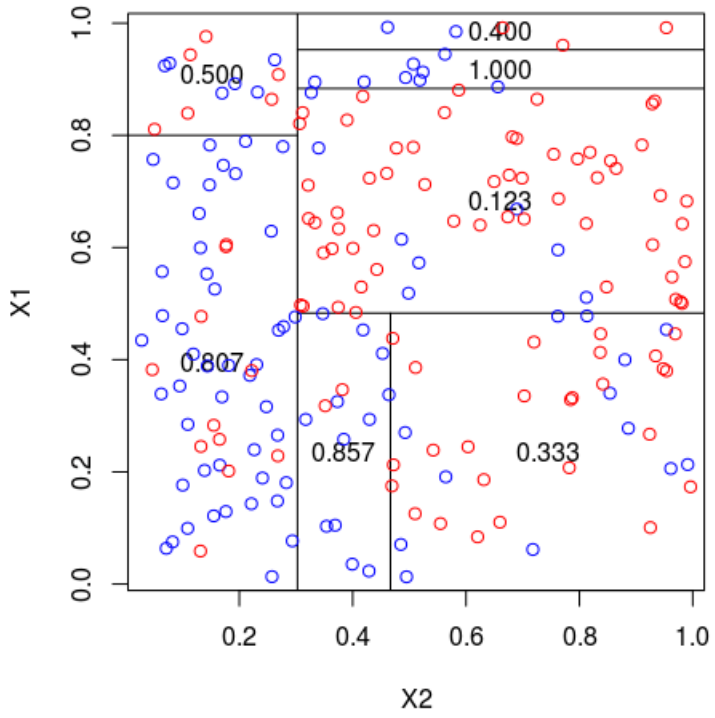
- **Decision Tree: main Idea**
  - Split the feature space into a set of rectangles, and inside each subspace fit a simple predictive model
- **Some interesting characteristics**
  - Recursive algorithm
  - Greedy construction
  - Automatic feature selection
  - A good way to start data exploration and analysis
- **Interpretability**
  - The deeper the tree, the harder is to analyse and reason
  - The variables near the root are the most ‘important’
  - Aim for depth = 3 for a good balance of performance and explanation to a large audience



Model-specific: Local

- **Decision propagation or rule unwrapping**
  - Input and output space example, Regression Tree
  - Two input variables, X1 and X2, and a single output variable $Y \in [0, 1]$



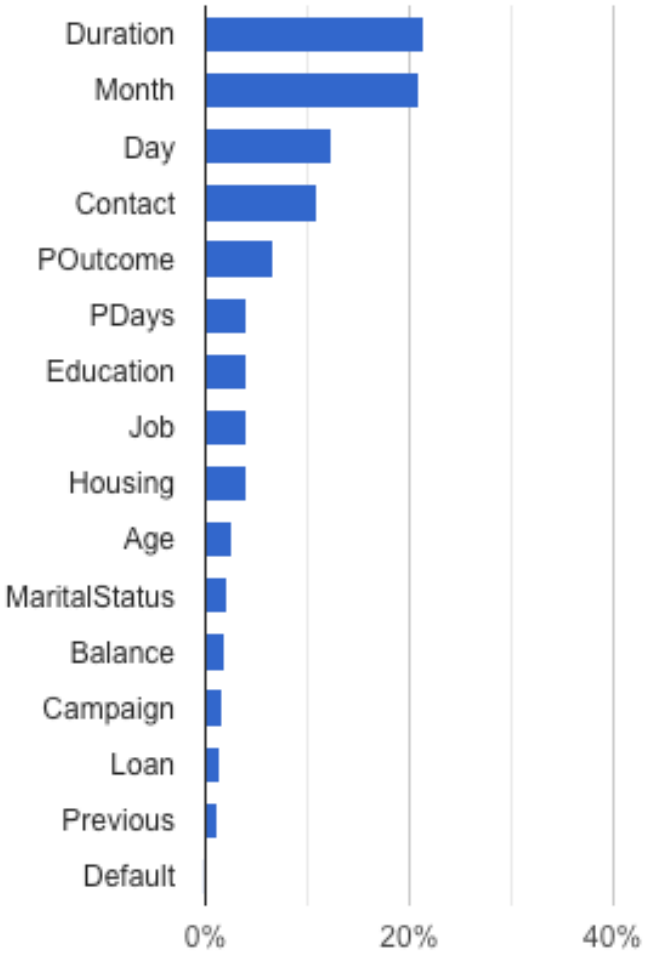


  - If Subject 1 has X1 = 0.6 and X2 = 0.2, then ?

Model-agnostic: Global

Feature Importance

- Often, the input predictor variables are seldom equally relevant. Often, only a few of them have substantial relevance
- **Feature Importance** metrics is a way in which “black-box” models become transparent
  - Weight of each feature
  - Positive/negative relationship
- **Some ways to calculate**
  - Weighted node impurity reduction
  - Out-of-bag error degradation via Row permutation
- **Caveat**: will underestimate “importance” in the presence of correlated variables



Model-agnostic: Global

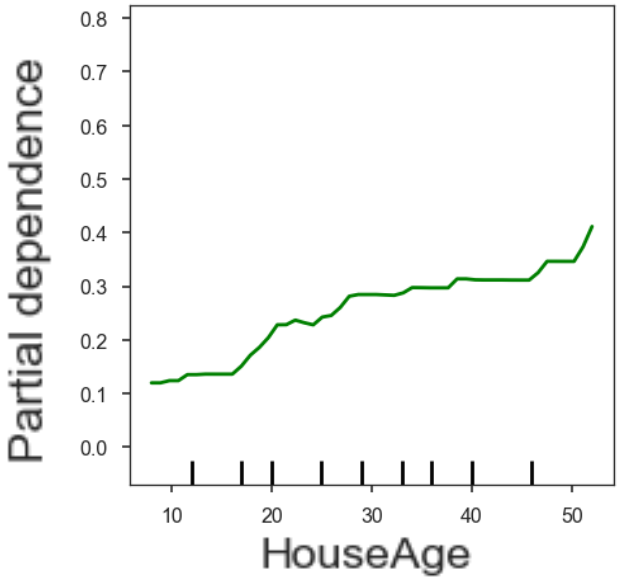
Partial Dependence

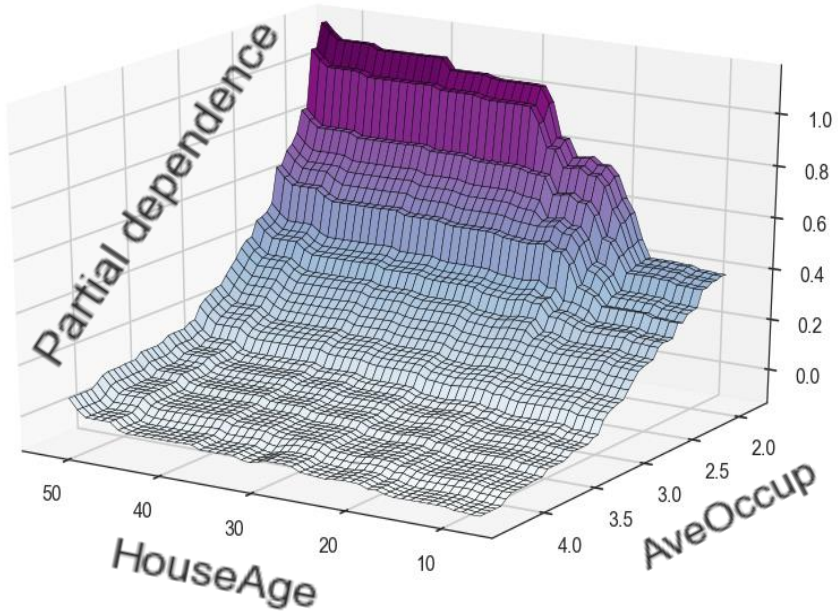
- Partial dependence functions can be used to interpret the results of any "black box" learning method.
- Given a feature set *S* (feature A, say) and its complement *C,* partial dependence can be estimated by:

$$\hat{y}_S(\boldsymbol{x}_S) = \frac{1}{n}\sum_{i=1}^{n} F(\boldsymbol{x}_S, x_{iC})$$

for every value $\boldsymbol{x}_S$ in the range

- For a given model, it involves a heavy amount of computation
- However, for decision trees it is rapidly computed from the tree itself without reference to the data

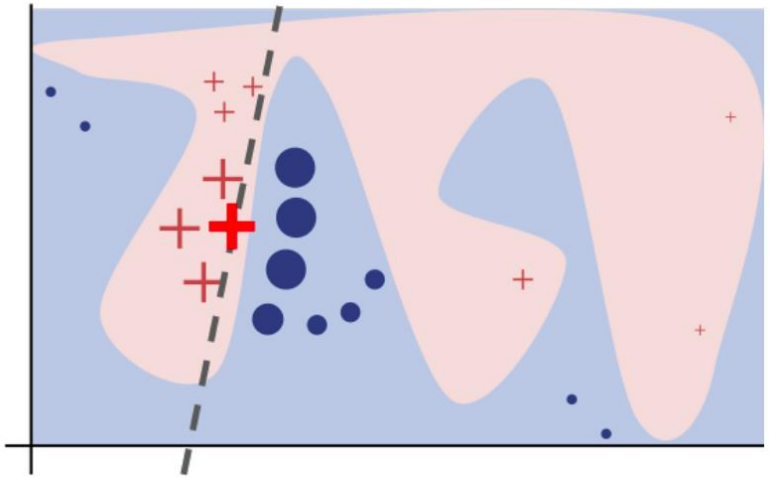




Model-agnostic: Local

LIME

- Local Interpretable Model-Agnostic Explanations (LIME)
  - Important research (2016), influencing many methods that came afterwards.
  - Provided some foundational thought, particularly about fidelity-interpretability trade-off
- What is to be Local for LIME
  - Giving more importance to samples at a vicinity of a data point x that we aim to provide a local explanation
- What is to be Interpretable for LIME
  - Belonging to a class of potentially interpretable models like decision trees, linear models, etc.
- What is to be Model-agnostic for LIME
  - Without knowledge of the underlying machine learning model, that is, if it is a Neural Network, SVM, etc.



Algorithm 1 Sparse Linear Explanations using LIME

Require: Classifier f, Number of samples N
Require: Instance x, and its interpretable version x'
Require: Similarity kernel π_x, Length of explanation K

$\mathcal{Z} \leftarrow \{\}$
for $i \in \{1, 2, 3, ..., N\}$ **do**
  $z'_i \leftarrow sample\_around(x')$
  $\mathcal{Z} \leftarrow \mathcal{Z} \cup \langle z'_i, f(z_i), \pi_x(z_i) \rangle$
end for
$w \leftarrow$ K-Lasso$(\mathcal{Z}, K)$ ▷ with z'_i as features, $f(z)$ as target
return w