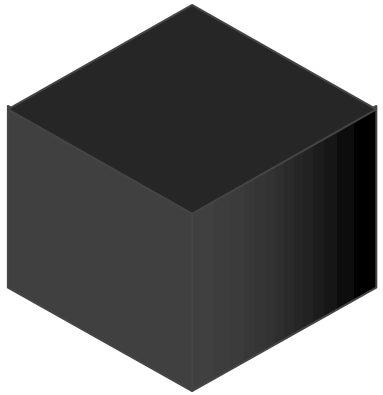
Explainable AI - Taxonomy

Explainable and Trustworthy AI

Eliana Pastor

1. Transparency & Explainability

Most AI models are black boxes and lacks interpretability



Opaqueness of a model can be at multiple levels, involving

- Data
- Model/System/Algorithm
- Learned function/pattern → reasons behind its functioning
- Intention and business model of the AI product

These elements should be transparent – clear, disclosed - to the end users

A definition of Explainability/Interpretability

The ability to explain or to present the **reasoning** behind the decisions of **AI systems** or its technical process in understandable terms **to humans**

Focus on humans as targets of explainability

Multiple terms – Not monolithic concept -

Many terms and definitions are adopted. Slight variations, overlapping concepts

No agreement within the ML community on the definition

- **Interpretability**: An interpretable model is **transparent** in its operation and provides information about the relationships between inputs and outputs
  - Focus more on models that are **inherently interpretable**
- **Explainability**: ability to explain the decision-making process of an AI model in terms understandable to the end user. Emphasize the ability to provide an **explanation of the decisions**.
  - Focus more on models that are by design incomprehensible to humans

Multiple terms – Not monolithic concept -

- **Understandability**: an interpretable model is a model that can be understood, in a reasonable amount of time
- **Comprehensibility**: emphasis as enabler to produce knowledge, as it is enabled if the pattern identified by AI systems are comprehensible
- **Intelligibility:** a model is intelligible if it is interpretable by humans
- **Mental fit**: ability for a human to grasp the model

These terms are often used interchangeably.

We will mostly use the terms interpretable and explainable. We will used them interchangeably, but emphasizing that interpretability mostly relate to interpretable by design models while explainability to explain models, making them interpretable

Desiderata of explainability research

By understanding an AI model and its predictions, we also achieve other desiderata – also requirements for Trustworthy AI

- Trust
- Fairness and ethical decision making
- Robustness
- Informativeness and knowledge

Desiderata of explainability - Trust

If we can understand the model, we can decide whether to trust it or not

Example Pneumonia risk case:

- The interpretable model learned that an history of asthmas is associated with lower chance of dying from pneumonia. Being interpretable, experts could analyze it and decide they could not trust it

Example of application to Apple Card:

- Users did not trusted the model and had fairness concerns since the model was opaque and could not understand the reasons behinds its functioning

Desiderata of explainability - Fairness

If we can understand the model, we can assess if it bases its functioning on sensitive and protected information or make decisions based on discriminatory aspects

Example COMPAS:

- Analysis on the model revealed its biased predictions

Example algorithms for predict outcomes in UK exams

- Being opaque, concerns on its fairness

Desiderata of explainability - Robustness

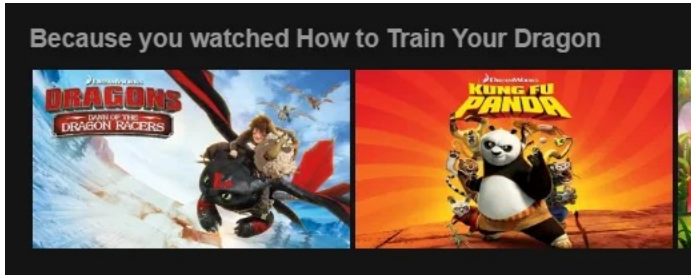
If we can inspect erroneous predictions, we can actively work on model debugging

"a young boy is holding a baseball bat."

Desiderata of explainability - Informativeness

If we can reveal the reasons behind model predictions, we inform users

Example recommentations



Example loan approval

‘We rejected your loan request because your *‘your income was insufficient or unstable’*”

Desiderata of explainability - Knowledge

If we can inspect model behavior, we can potentially gain new form of knowledge

Example AlphaGo
“It’s not a human move. I’ve never seen a human play this move,” says European Go champion Fan Hui. “So beautiful.”

AlphaGo

Source

A taxonomy of Explainable AI

At which step of the ML Pipeline?

Pre-modeling, modelling, post-modelling explainability

Is the explanation method general?

Model dependent vs model agnostic

What do we explain?

The global model, subgroups, a local prediction

How do we represent the explanations?

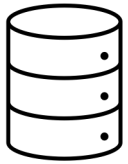
Feature importance, rules, visualization-based explanation, explanation by example

How can we derive explanations?

Explaining by removing, local surrogates, gradient-based

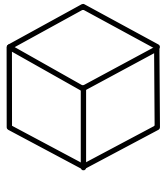
Stages of Explainability

Explainability involves the entire AI development pipeline



Pre-modelling explainability

Before building the model
- Data exploration
- Data selection
- Feature engineering



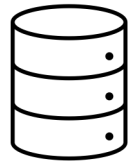
Explainable modeling

Build inherently interpretable models
- Manage the accuracy and interpretability trade-off

Post-modelling explainability

After model development
- Explaining predictions and behavior of trained models

Stages of Explainability – Pre-modelling explainability



Before the actual modeling process

The goal is to gain understanding of the data and preprocess them for the following model development preserving their understandability. It may also involve assessing potential biases in the data and addressing them before modeling.

It includes:

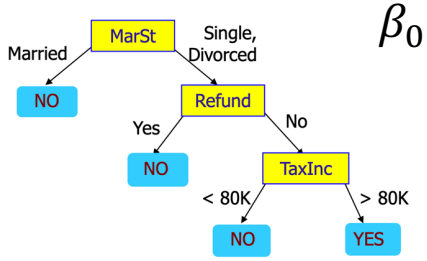
- Exploratory data analysis: extract a summary of the main characteristics of a dataset
- Interpretable feature engineering: selection and preprocessing of features preserving their interpretability
- Data description and summarization: standardize documentation, ensure proper communication between data providers and users of datasets

Stages of Explainability – Explainable modelling

Design, train and adopt more interpretable/explainable models

- Adopting **an inherently explainable models**
  - does not automatically guarantee explainability (e.g., deep trees, linear models on high dimensional data)
  - Problem of explainability vs performance trade-off: interpretable models are typically less performing

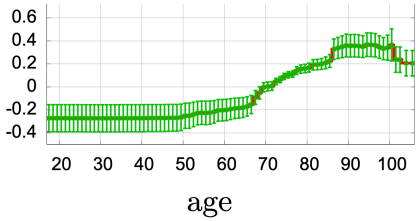
Trees



Linear models

$$\beta_0 + \sum_i \beta_i x_i$$

GAMs, GA²Ms, GLMs



Caruana et al. "Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission." KDD 2015

Decision sets - Rules

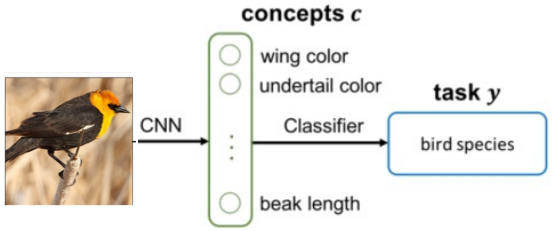
If Respiratory-Illness=Yes **and** Smoker=Yes **and** Age≥ 50 **then** Lung Cancer

If Risk-LungCancer=Yes **and** Blood-Pressure≥ 0.3 **then** Lung Cancer

If Risk-Depression=Yes **and** Past-Depression=Yes **then** Depression

Lakkaraju et al. "Interpretable decision sets: A joint framework for description and prediction." KDD 2016

Concept-based models



Koh, Pang Wei, et al. "Concept bottleneck models." ICML 2020.

Stages of Explainability – Explainable modelling

- **Targeting interpretability by design**
  - Design high-performing models imposing interpretability constraints to enable their interpretability
  - e.g., Explainability via regularization
    - Apply regularization to improve model explainability

$$\min_{f \in F} \sum_i Loss(f, x_i, y_i) + \text{InterpretabilityPenalty}(f),$$

$$\text{subject to Interpretability constraint}(f)$$

Examples of constraints: number of leaves of a decision tree, weights different than 0 for a linear models

Problem: these models could still underperform compared to more complex models

Cynthia Rudin et al. “Interpretable machine learning: Fundamental principles and 10 grand challenges”, 2021

Stages of Explainability – Explainable modelling

- **Explanations-in-the-loop**

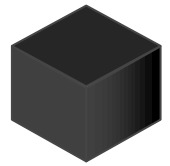
  Train AI systems **to** jointly **provide** a prediction and its **explanation**

  - e.g., Teaching Explanations for Decisions (TED) framework:
    - train the task model by also providing user rationales for the ground truth labels, i.e., ground truth explanation.
    - At deployment time, the system provides the outcome and its corresponding explanation

- Drawbacks:
  - Require a dataset annotated with explanations
  - Explanations may not necessarily reflect of how model predictions were made but what humans expects
  - Faithfulness to the model vs Plausibility

Hind et al. TED: Teaching AI to Explain its Decisions. AIES 2019

Stages of Explainability – Post-modelling explainability



After the modeling process

Generate explanations for pre-trained models

- Most high-performing models are black boxes – optimizing only performance
- The majority of XAI approaches focus on post-modelling explainability as they address interpretability concerns for pre-trained black box models
    - Also known as post-hoc explainability methods.

- These approaches highly differentiate for the other dimensions of XAI taxonomy: Explainability generality, Explainability scope, Explanation representation, Methods to derive explanations

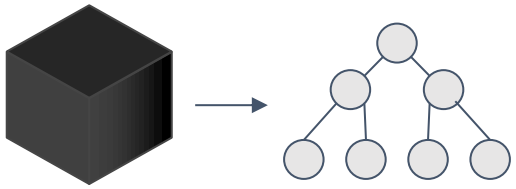
Generalizability of Explainability

- **Model dependent solutions**
  - Only applicable for specific models
    - e.g., specific approaches for explaining SVM, approaches for explaining a specific neural network
  - Relies on the model structure/properties

- **Model agnostic solutions**
  - Applicable to any model
  - Relies on the model as an oracle (model predictions, output probabilities)

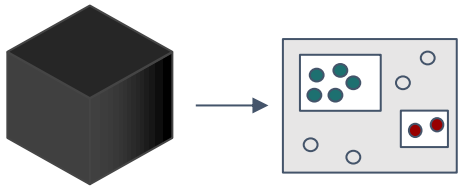
Scope of Explainability

What do we explain?



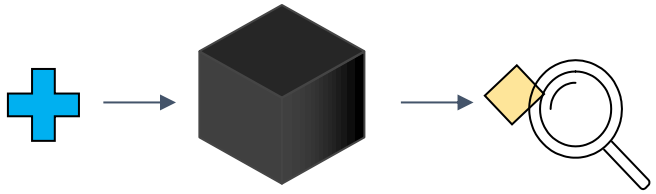
Global

How the model globally works



Subgroup

How the model behaves in data subgroups



Individual/local

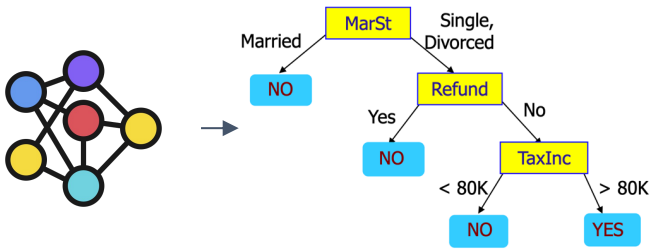
Explaining the reasons behind individual predictions

Scope of Explainability – Global explanations

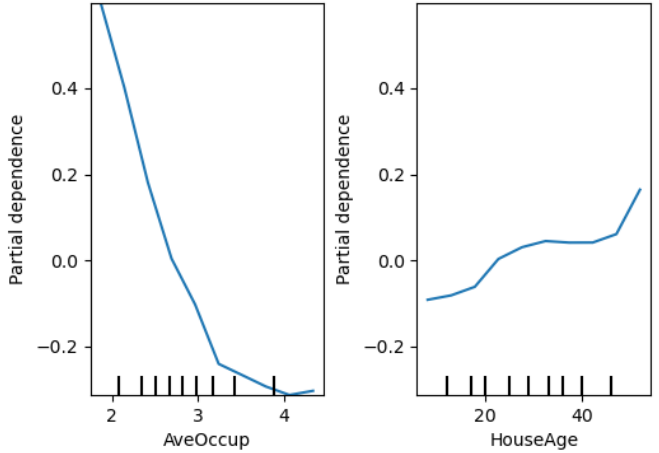
Global methods describe the overall behavior of model

Explain how the model works in general

Interpretable global surrogate models

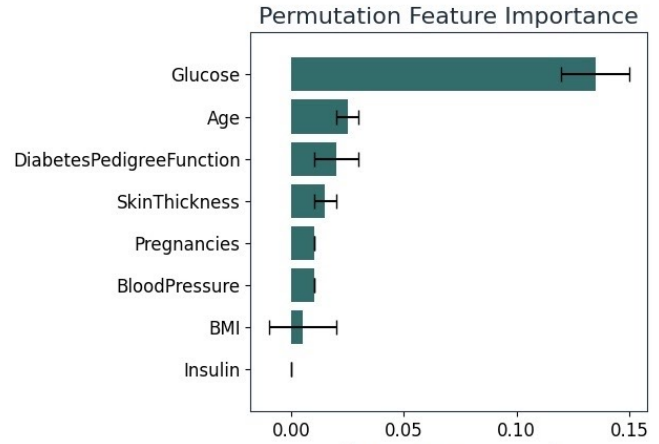


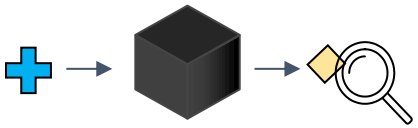
Partial dependence plots



Dependence between the target response and an input feature of interest

Permutation feature importance.





Scope of Explainability – Individual/Local explanations

Local explainability methods explain individual predictions

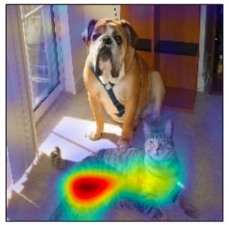
Why does the model make this decision?

- Explain a prediction is a simpler task than explaining an entire model
  - Simple to approximate the model behavior for a single instance
- A single local explanations is easier to understand and analyze compared to a global explanation

Multiple approaches have been proposed to generate local explanations. They mainly differentiate on how explanations are represented and how to generate explanations.

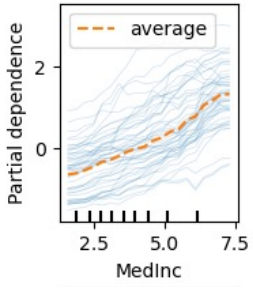
Explanation representation

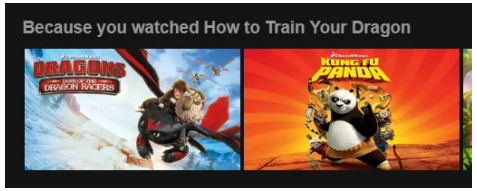
- Feature importance
- Local rules
- Visualizations
- Explanations by example



(c) Grad-CAM ‘Cat’

If Country is US, married, work hours > 45
→ Income>50K





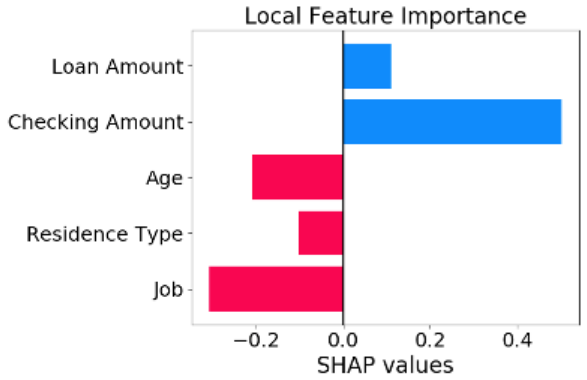
Explanation representation

Feature importance – input attribution

Indicate how much each feature contributed to the prediction for a given instance.

- These attributions can be presented numerically, graphically, or in a tabular format to show which features are most influential in driving the model's predictions.

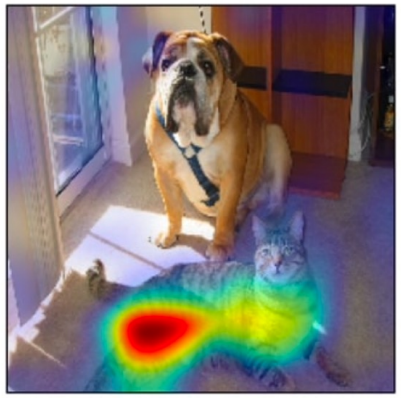
Structured data



Text

I am really happy

Images



(c) Grad-CAM ‘Cat’

Explanation representation

Rule explanations

Local rules that provide insights into how a model behaves for a specific instance

- Often extracted from local surrogate models.
- They can be presented as logical statements, decision paths, or decision rules that describe the model's behavior for a particular input.

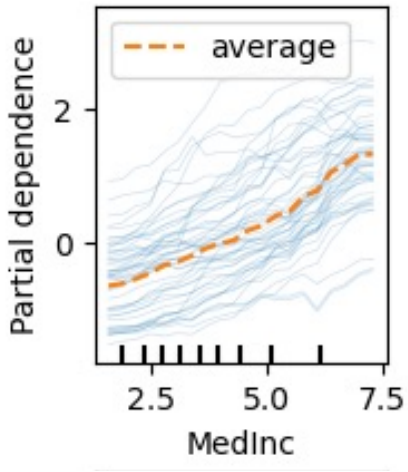
If Country is US, married, work hours > 45
→ Income>50K

Explanation representation

Visualization-based explanations

the use of visual representations to explain the behavior

Individual Conditional Expectation (ICE) plot



Alex Goldstein et al. “Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation”. In: Journal of Computational and Graphical Statistics 24.1 (2015)

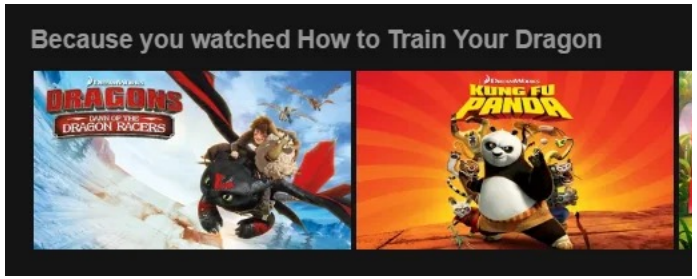
Explanation representation

Explanations by example

Example-based explanation methods select or generate instances to explain

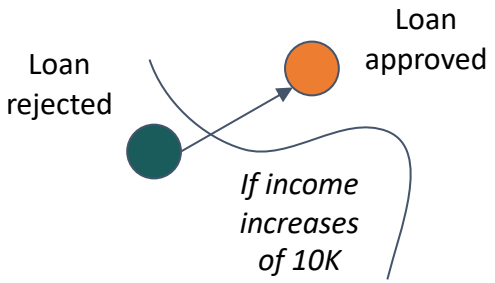
Prototypes

Representative instance(s) of prediction to explain



Counterfactual

Slightest change that changes the prediction



Adversarial examples*

Counterfactuals where the small changes are applied to fool the model rather than interpret it

*Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572 (2014).

Methodology to derive explanations

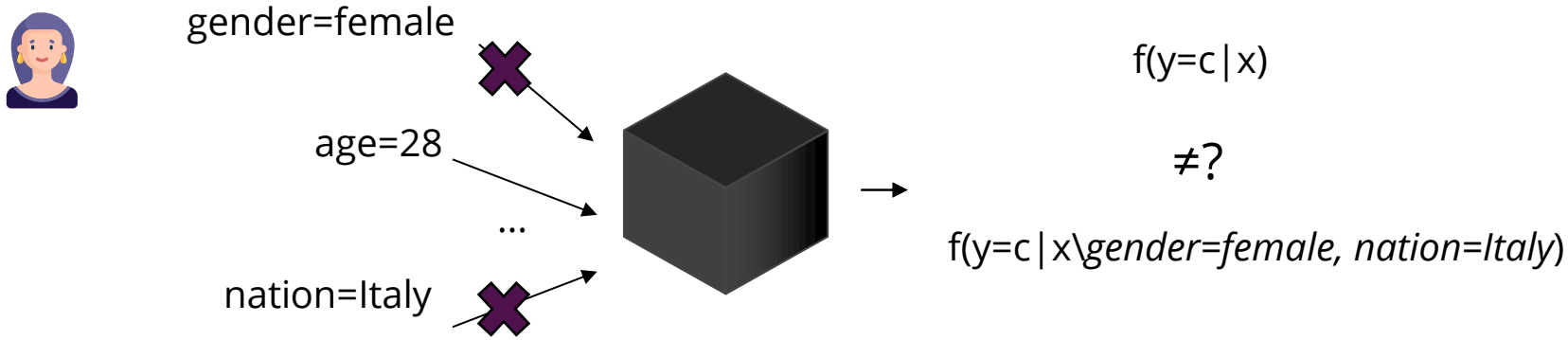
- Explaning by removing
- Local surrogate interpretable model
- Gradient-based explanation methods
- Counterfactual methods

Methodology to derive explanations

Explaining by removing

Also known as *occlusion-based* or *perturbation-based* since often the occlusion is performed via perturbations

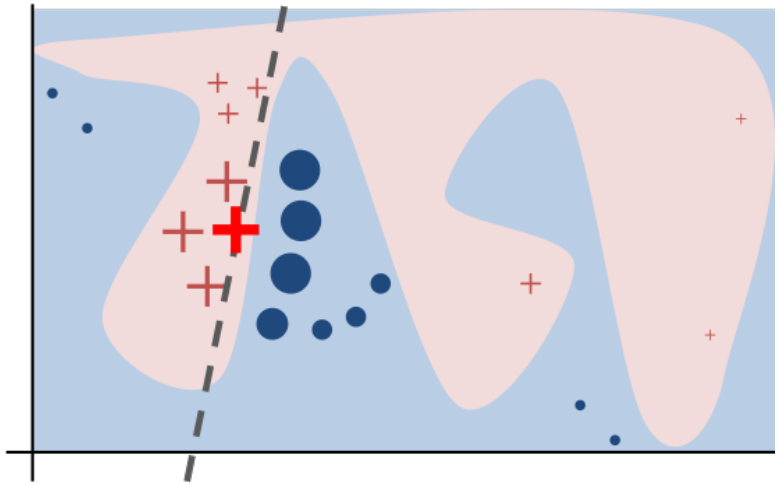
The idea is to remove one or more input feature (or simulate the removal) to quantify the feature influence



Methodology to derive explanations

Local surrogate interpretable model

- Approximate the behavior of complex black-box models in the locality of a prediction
- Involve training interpretable models, such as decision trees or linear models, on data points in the vicinity of the instance to be explained
- Provide insights into how the black-box model behaves in the neighborhood of a specific data point

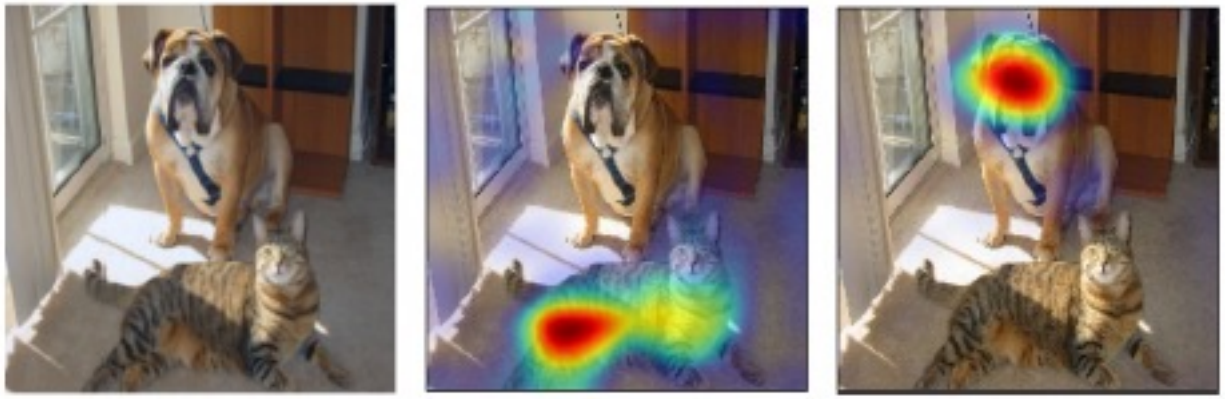


[1] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "'Why should i trust you?' Explaining the predictions of any classifier." KDD 2016
[2] Guidotti, Riccardo, et al. "Local rule-based explanations of black box decision systems." 2018

Methodology to derive explanations

Gradient-based explanation methods

Leverage gradients of the model's output with respect to input features to provide insights into how changes in input features affect the model's predictions.

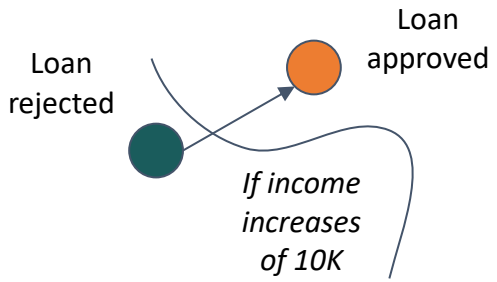


Original Image | Grad-CAM 'Cat' | Grad-CAM 'Dog'

Methodology to derive explanations

Counterfactual methods

- Generating alternative instances (counterfactuals) with similar characteristics but different predictions, to understand how small changes in input features affect the model's output



References

- Adrien Bibal and Benot Frenay. Interpretability of Machine Learning Models and Representations: an Introduction,
- Nakhaeizadeh, Gholamreza, and Alexander Schnabl. "Development of Multi-Criteria Metrics for Evaluation of Data Mining Algorithms." *KDD*. 1997.
- Bibal, Adrien, and Benoît Frénay. "Interpretability of machine learning models and representations: an introduction." *24th european symposium on artificial neural networks, computational intelligence and machine learning*. CIACO, 2016.
- Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission KDD 2015
- Lipton, Zachary C. "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." Queue 16.3 (2018): 31-57.
- Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." KDD 2016.
- Bahador Khaleghi. An Explanation of What, Why, and how of Explainable AI (XAI) Toronto Machine Learning Summit in November
- Bahador Khaleghi. The How of Explainable AI: Pre-modelling Explainability
- https://christophm.github.io/interpretable-ml-book/