Why & How Machine Learning Models should explain themselves

Machine Learning Interpretability

@XavierRenard | @detyniecki
AXA / GO / REV / Research & Development
rev was built to make AXA a tech-led company

OUR MISSION

AXA rev (Research, Engineering, & Vision) explores and scales the value of data and emerging technologies with the potential to disrupt the current insurance business model and to shape future opportunities in order to be a better partner in our customer’s lives.
Can A.I. Be Taught to Explain Itself?

As machine learning becomes more powerful, the field’s researchers increasingly find themselves unable to account for what their algorithms know — or how they know it.

By CLIFF KUANG

Artificial Intelligence’s ‘Black Box’ Is Nothing to Fear

By Vicky Pivalle

Ethique et intelligence artificielle : récit d’une prise de conscience

Is Explainability Enough? Why We Need Understandable AI

co-authored with McCre Lake, Talent & Organization Strategy Senior Manager atAccenture

Artificial Intelligence is quietly becoming ubiquitous in personal and professional lives in ways we both observe and others we don’t see at all. Artificial Intelligence is used to influence life-changing decisions, such as whether or not you get hired to that dream job, who you will date, and whether or not you’ll be approved for a loan for your first home. However, we have little insight into how critical decisions are made with AI. As a result, there is increasing demand (and legislation) to ensure the influence of those technologies is understood.
Machine Learning Interpretability Impacts the Business

**Improve Model’s Quality**
- Improve models, features, robustness, fairness, etc.
- Identify data leakage & data drift
- *e.g.* Understand origin of wrong predictions

**Reassure Users & Business Owners**
- Trust by explanation: improve ML acceptance
- Help to take ML prediction-based decision
- *e.g.* Assess reasonable behaviour if deployment

Use-Case in Fraud: Analysts insist to understand why there is an alert

**Law & Ethics compliance**
- Right to explanation
- Assess model’s fairness
- Inform customers

**Gain Knowledge on Business’ processes**
- Insight of revenues or value-generating application
- *e.g.* Credit scoring, fraud detection, etc.
Machine Learning Interpretability
APPLIED TO AXA’S HEADQUARTERS

Classification
With InceptionV3:

Most probable labels:
- Building
- Minivan
- Traffic light
Machine Learning Interpretability
APPLIED TO AXA’S HEADQUARTERS

Classification (With InceptionV3)

Class label: Building
+ building structure, windows
- cars

Class label: minivan
+ minivan

Class label: traffic light
+ cars & yellow lights
Machine Learning Interpretability
EVALUATE MACHINE LEARNING MODELS BEYOND ACCURACY SCORES

- What has been learned by the model?
- Why a particular prediction has been made?
- Where is the model {correct; wrong}?
- What can be done to change the prediction?
- Usually aggregated accuracy score

- Is the model robust?
- How does the model behave in areas with few data?
- Is the model fair?
- How the prediction is affected by small changes in input?

Description of the problem to solve
Tabular data, unstructured data, etc.

Machine Learning Model
Prediction / Decision
Trade-off Interpretability-Accuracy

**Accurate** Machine Learning Models are **not Interpretable** (usually)

**Simple** machine learning model
- e.g. Decision Tree

**Blackbox** machine learning model
- e.g. Random Forest, CNN (Inception…)

```
Status_of_existing_checking_account <= -0.07
  gini = 0.423
  samples = 700
  value = [[213, 487]
            [487, 213]]

  True

  gini = 0.494
  samples = 379
  value = [[168, 211]
            [211, 168]]

  False

  gini = 0.241
  samples = 321
  value = [[45, 276]
            [276, 45]]
```
Trade-off Interpretability-Accuracy

**Accurate** Machine Learning Models are not **Interpretable** (usually)

**Simple** machine learning model

e.g. Decision Tree

- Status_of_existing_checking_account <= -0.07
  - gini = 0.423
  - samples = 700
  - value = [[13, 487]
    [487, 243]]

  **True**

- gini = 0.494
  - samples = 379
  - value = [[168, 211]
    [211, 168]]

  **False**

- gini = 0.241
  - samples = 321
  - value = [[45, 276]
    [276, 45]]

**Blackbox** machine learning model

e.g. Random Forest, CNN (Inception…)

**Decision**: credit or not

One **path** → simple **explanation**
Trade-off Interpretability-Accuracy

**Accurate** Machine Learning Models are **not Interpretable** (usually)

- **Simple** machine learning model
e.g. Decision Tree
- **Blackbox** machine learning model
e.g. Random Forest, CNN (Inception…)

**Decision**:
- credit or not

One path → simple explanation

**Final decision**: aggregation of each decision

**Explanation**: no consensus
Trade-off Interpretability-Accuracy

Accurate Machine Learning Models are not Interpretable (usually)

**Simple** machine learning model
e.g. Decision Tree
→ Interpretable
→ Less accurate

**Blackbox** machine learning model
e.g. Random Forest, CNN (Inception…)
→ Uninterpretable
→ More Accurate

Status of existing checking account $\leq -0.07$
- gini = 0.423
- samples = 700
- value = [[413, 487], [487, 213]]

Decision: credit or not

One path → simple explanation

One path → One decision by base model
Final decision: aggregation of each decision

Explanation: no consensus
Taxonomy of Interpretability Approaches

Interpretable Model
- Decision tree, Linear model

Post-Hoc Model Specific
- Specific feature importance extraction
  (e.g. feature’s gini contribution for random forest)

Post-Hoc Model Agnostic
- Prototype Selection
- Sensitivity Analysis
  - PDP, ICE

Surrogate Model

Local Model
- LIME, SHAP, Shapley values, Anchor, LAD, LS, etc.

Global Model
- Trepan, SLIMs, GAM, etc.
Locality Issue (2018 ICML WHI)

A widely used approach -LIME- is inaccurate
Locality Issue \((2018\ ICML\ WHI)\)

Our proposition: find the frontier first

Better black-box frontier approximation → more accurate explanations
Gather Related Variables for More Interpretable (Surrogate) Decision Trees

- Global & Local explanation of a black-box classifier based on a decision tree and concepts
  - In the presence of correlated variables
  - Or expert-defined groups of variables

**Figure 1.** Concept Tree trained on FRED-MD macroeconomic dataset. Variables are grouped by Concepts to constrain the training of an interpretable surrogate decision tree.
Unjustified Counterfactual Explanations (2019 IJCAI)  
The Dangers of Post-hoc Interpretability: Unjustified Counterfactual Explanations

- Instance close to the original observation predicted in a different class
  - They can be a consequence of an artifact of the classifier
  - Unjustified by ground truth (training data)
  - Lack of robustness of the classifier / ood prediction
- To be justified a counterfactual example should be continuously connected to an instance of the training set
- Assessment procedure proposed
- Counterfactual explanation methods vulnerable to unjustified counterfactual examples
Imperceptible Adversarial Attacks on Tabular Data (on going)
ML interpretability * Adversarial ML

Figure 1: Intuition of adversarial samples generation using feature-importance, in black, compared to traditional methods in gray. We observe that feature-importance based perturbations follow only the Feature 2 directions.
Why & How Machine Learning Models should explain themselves

Machine Learning Interpretability

@XavierRenard | @detyniecki
AXA / GO / REV / Research & Development