Global Explainability (XAI) Techniques

Quantifying the uncertainty of the explanations

Vasilis Gkolemis[1]

[1] ATHENA Research and Innovation Center

November 2021

Program

1. Intro to XAI
2. Feature Effect
3. Feature Interaction
4. Heterogeneous effects
5. Feature Importance
6. Extras

Program

1. Intro to XAI
2. **Feature Effect**
3. Feature Interaction
4. Heterogeneous effects
5. Feature Importance
6. Extras

Program

1. Intro to XAI
2. Feature Effect
3. Feature Interaction
4. Heterogeneous effects
5. Feature Importance
6. Extras

Feature Interaction - Motivation

- Is Feature Effect a good approach?
  - Interpretability? very good, easy intuition
  - Fidelity? it depends..
- Additive case: $f(\boldsymbol{x}) = f_1(x_1) + f_2(x_2)$
  - Generalized Additive Models
  - X-by-design
- Non-additive case: $f(\boldsymbol{x}) = f_1(x_1) + f_2(x_2) + \underbrace{f_{12}(x_1, x_2)}_{\textit{interaction}}$
  - how to distribute $f_{12}(x_1, x_2)$ to x_1 and x_2?
  - Research question! Later: uncertainty could help
- f is unknonw, so first, someone must inform about the interaction terms
- Feature Interaction Methods!

Problem Statement

When features interact with each other in a prediction model, the prediction cannot be expressed as the sum of the feature effects, because the effect of one feature depends on the value of the other feature. Aristotle's predicate "The whole is greater than the sum of its parts" applies in the presence of interactions.[1]

[1] Interpretable Machine Learning book

H-statistic

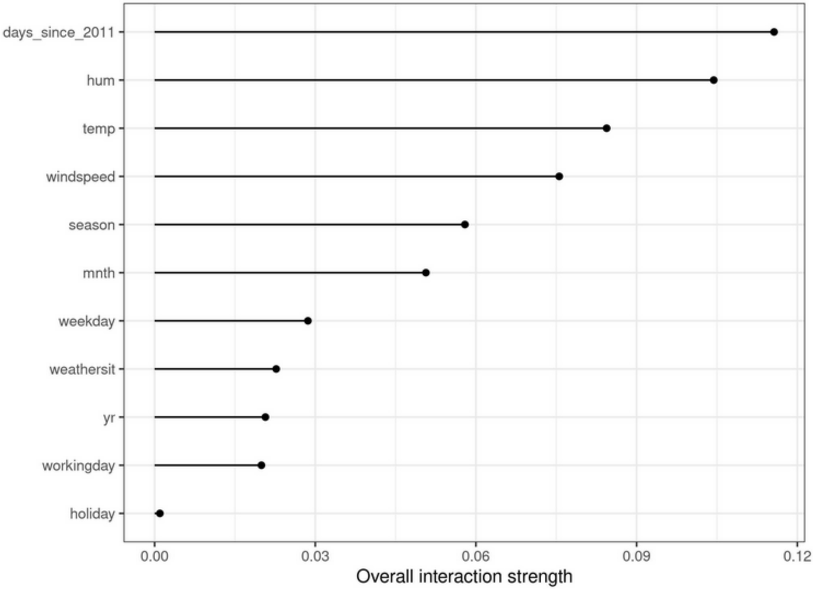
- Level of interaction between feature *i* and feature *j*

$$\mathcal{H}^2_{jk} = \frac{\sum_{i=1}^{n} \left(PD_{jk}(x_j^{(i)}, x_k^{(i)}) - PD_j(x_j^{(i)}) - PD_k(x_k^{(i)}) \right)^2}{\sum_{i=1}^{n} PD^2_{jk}(x_j^{(i)}, x_k^{(i)})}$$

- Level of interaction between feature *i* and all the other features

$$\mathcal{H}^2_{j} = \frac{\sum_{i=1}^{n} \left(f(x^{(i)}) - PD_j(x_j^{(i)}) - PD_{-j}(x_{-j}^{(i)}) \right)^2}{\sum_{i=1}^{n} f^2(x^{(i)})}$$

H-statistic


days_since_2011
hum
temp
windspeed
season
mnth
weekday
weathersit
yr
workingday
holiday
0.00
0.03
0.06
0.09
0.12
Overall interaction strength

Other approaches

- Greenwell's interaction index
  - PDP-based method
  - A Simple and Effective Model-Based Variable Importance Measure
- SHAP interaction index
  - SHAP-based method
  - Consistent Individualized Feature Attribution for Tree Ensembles

Program

1. Intro to XAI
2. Feature Effect
3. Feature Interaction
4. Heterogeneous effects
5. Feature Importance
6. Extras

Interaction implies heterogeneity

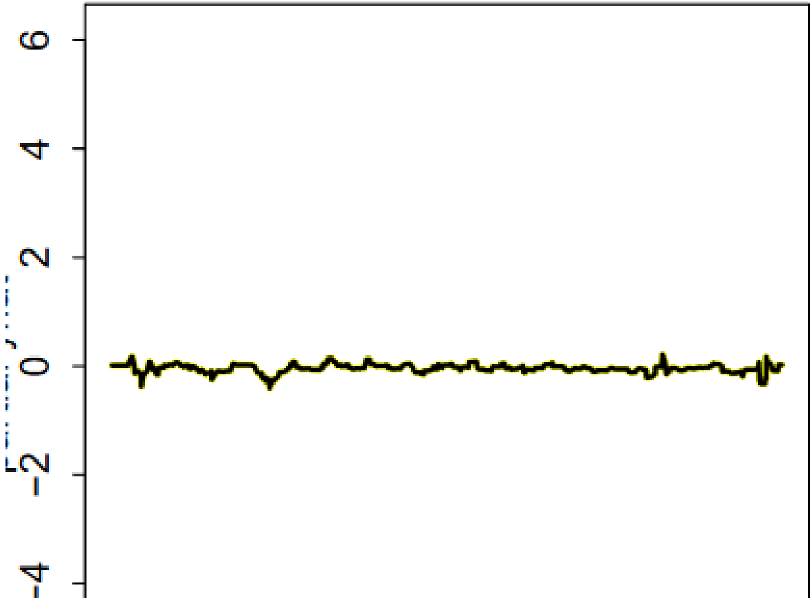
The interaction index measures the contribution of the interaction terms:

- $f(x_1, x_2) = x_1^2 + \log(x_2) + \alpha x_1 x_2^3$
- $\alpha = 0.1 \rightarrow$ low interaction index $\rightarrow$ high fidelity
- $\alpha = 100 \rightarrow$ high interaction index $\rightarrow$ low fidelity

---

But it does not say how the interaction terms influence the feature effect plots

Example



Program

1. Intro to XAI
2. Feature Effect
3. Feature Interaction
4. Heterogeneous effects
5. **Feature Importance**
6. Extras

Program

1. Intro to XAI
2. Feature Effect
3. Feature Interaction
4. Heterogeneous effects
5. Feature Importance
6. Extras