Fast and accurate regional effect plots for inspecting black-box models fit on tabular data

Vasilis Gkolemis[1,2], Christos Diou[1], Eirini Ntoutsi[3], Theodore Dalamagas[2], Bernd Bischl[4], Julia Herbinger[4] and Giuseppe Casalicchio[4]

[1]Harokopio University of Athens
[2]ATHENA Research Center
[3]University of the Bundeswehr Munich
[4]Munich Center for Machine Learning (MCML), Department of Statistics, LMU Munich

Abstract
Feature effect methods are a novel approach to extract insights from tabular data with the following procedure. A black-box machine learning model is trained on a tabular dataset, a feature effect method explains its learnings, and the explanations (output of the feature effect method) are used to understand the data and and support decision making. Feature effect methods can be global or regional. Regional effects have the benefit of automatically identifying important subgroups but in the cost of added complexity. Global effects provide one explanation per feature, for example, how age influences the annual income, while regional effects give separate explanations based on subgroups, for example, how age influences the annual income differently for men and women. Existing regional effect methods suffer from efficiency and accuracy limitations. In this paper, we introduce rRHALE (regional RHALE), a method that overcomes these issues. rRHALE is notably more efficient, making it suitable for large datasets and complex models. It also handles correlated features effectively. We demonstrate its effectiveness with synthetic examples, comparing it to other feature effect methods, and apply rRHALE to a real-world scenario. The supporting code for this publication is available here.

Keywords
Explainability, Interpretability, Feature Effect, Regional Effect, Global Explanations

1. Introduction

There's been a significant increase in available data recently, prompting a need for data analytics tools to derive insights. Alongside this surge, Machine Learning methods have gained traction for their ability to automatically learn patterns and perform tasks like predictions. However, many Machine Learning models operate as black-boxes, meaning they take inputs and produce outputs without transparent inner workings. To address this, explainability techniques have emerged to shed light on these models' inner mechanisms. A promising approach to understanding data involves using machine learning models and then explaining them, essentially using model understanding to achieve data understanding.

- implements well-established global and regional effect methods, accompanied by an intuitive way to visualize the heterogeneity of each plot.
- follows a consistent and modular software design. Existing methods share a common API and novel methods can be easily added and compared to existing ones.
- demonstrates through tutorials the use of regional effects in real and synthetic datasets.

The 2nd World Conference on eXplainable Artificial Intelligence, July 17–19, 2024, Malta, Valetta

2. Regional RHALE

2.1. Method Description

2.2. Computational Complexity

All regional effect methods follow the CART-based approach of Algorithm(add reference), with computational complexity of ... Algorithm(cite) executes the prediction function $(D-1) \times N_2$ for computing the best split on a particular level, leading to a total of $(D-1) \times N_2 \times M$ execution of f.

3. Synthetic Example 1

The purpose of this example is to shwocase the superiority of Regional RHALE compared to Regional PDP, when features are correlated. In such cases, Regional PDP may erroneously find subregions due to out of distribution sampling.

Consider a black-box function $y = 3x_1 I_{x_3>0} - 3x_1 I_{x_3 \leq 0} + x_3$ and two different setting for the data generating distribution. In the non-correlated setting, all variables are uniformly distributed, i.e., $x_i \sim \mathcal{U}(-1, 1)$. In the correlated setting, we keep the same distributions for x_1 and x_2, but we set $x_3 = x_1$. We will focus on the effect (global and regional) of x_1 on y.

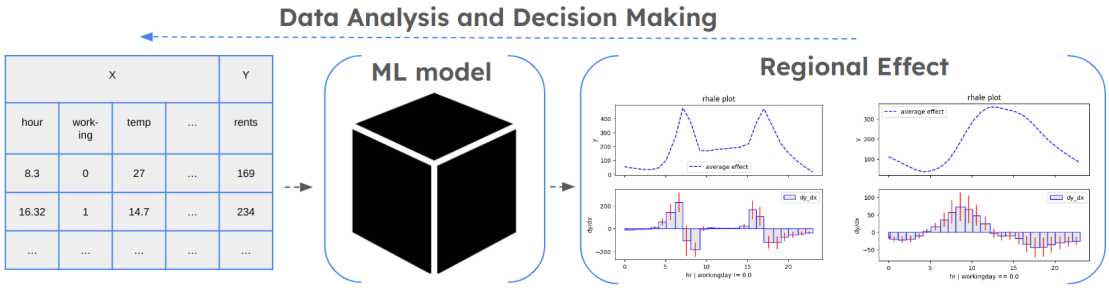
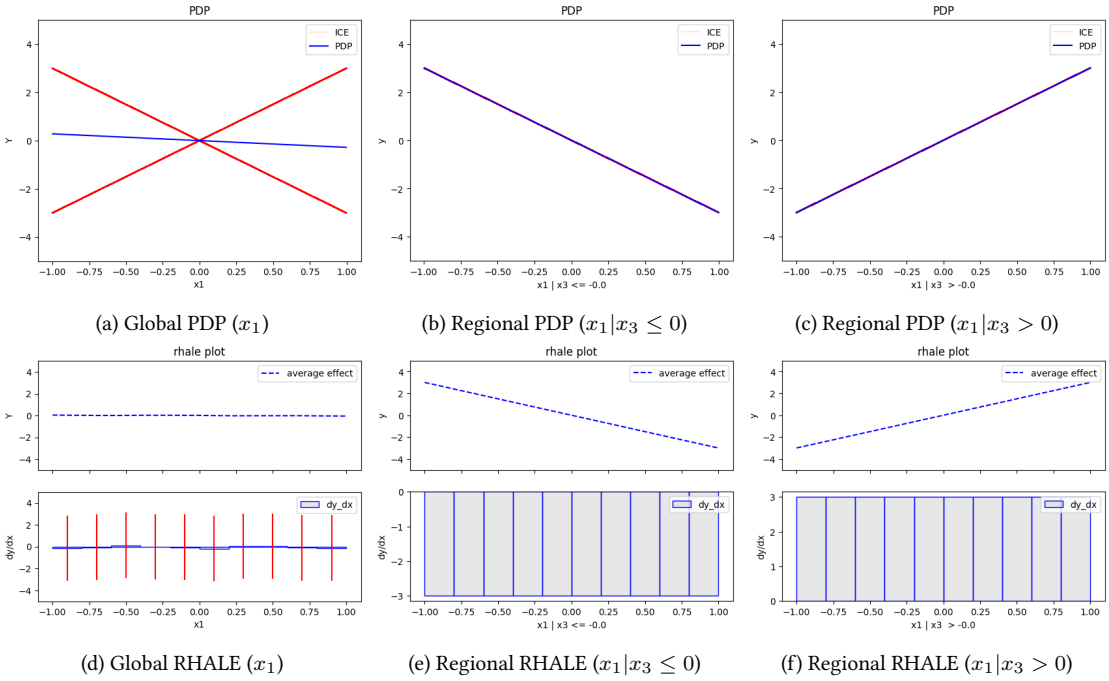


Figure 1: Data analysis and decision making pipeline: Utilizing regional effect plots to extract insights from tabular data.



(a) Global PDP (x_1) (b) Regional PDP ($x_1|x_3 \leq 0$) (c) Regional PDP ($x_1|x_3 > 0$)

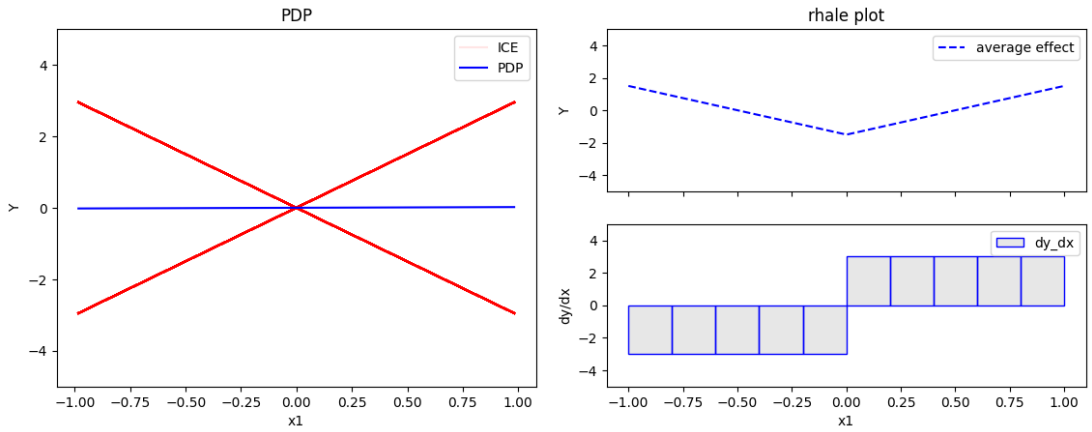
(d) Global RHALE (x_1) (e) Regional RHALE ($x_1|x_3 \leq 0$) (f) Regional RHALE ($x_1|x_3 > 0$)

Figure 2: Global and regional effect for the uncorrelated setting of synthetic example 1, using PDP and RHALE methods.

Non-correlated setting. The effect of x_1 is provoked by the interaction terms, i.e., $3x_1 I_{x_3>0}$ and $3x_1 I_{x_3\leq 0}$. The global effect will be $3x_1$ when $x_3 > 0$ (half of the times considering that $x_3 \sim \mathcal{U}(-1, 1)$) and $-3x_1$ when $x_3 \leq 0$ (the other half). This results in a zero global effect with high heterogeneity. If splitting into two subregions, $x_3 > 0$ and $x_3 \leq 0$, we get two regional effects, $3x_1$ and $-3x_1$, with zero heterogeneity each. In figure 2, we observe that both rPDP and rRHALE find correctly both global and regional effect.

Correlated setting. In the correlated case, due to $x_3 = x_1$, the interaction terms can be written as $3x_1 I_{x_1>0}$ and $-3x_1 I_{x_1\leq 0}$. This is because when $x_1 > 0$, $x_3 > 0$, so only the term $3x_1$ is active. Similarly, when $x_1 \leq 0$, $x_3 \leq 0$, making the term $-3x_1$ active.

4. Conclusion and Future Work



(a) Global effect

(b) Global effect

Figure 3: Global and interaction effects.