Plots for XAI: FANOVA Graph and FaithShapGraph

Lara Kuhlmann de Canaviri, Sonja Kuhnt

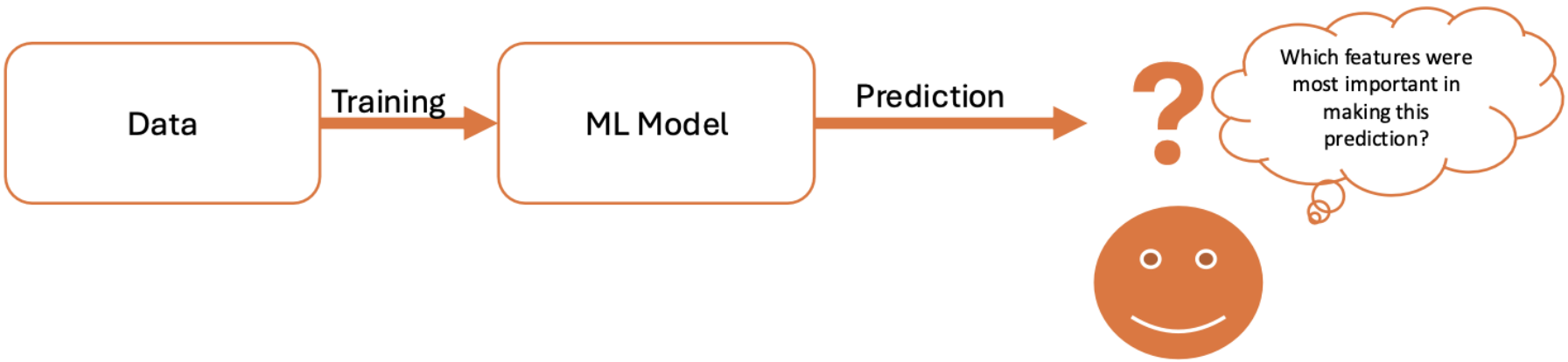
ENBIS-25 Piraeus

Transregio
391

Fachhochschule
Dortmund
University of Applied Sciences and Arts

AI

Modern AI systems usually operate as **black boxes**, making it difficult to understand how decisions are made.

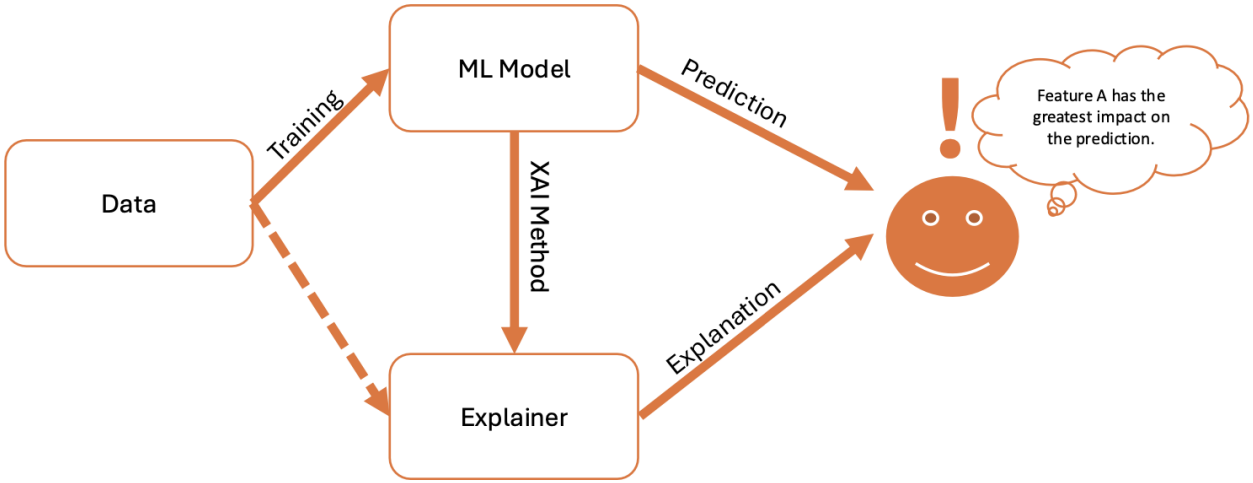


Lack of transparency is a major challenge for trustworthy AI.

Ribeiro, Singh, and Guestrin 2016

Explainable AI (XAI)

Explainable AI (XAI) provides transparent, interpretable insights into model predictions.



Typical approach:

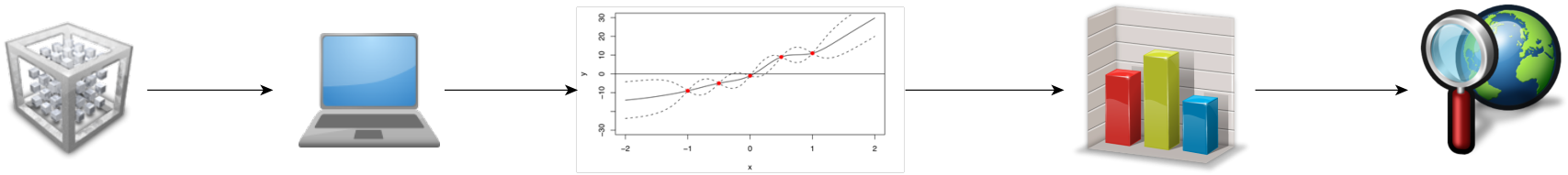
- Global surrogate models (e.g., interpretable decision trees)
- Local interpretable model-agnostic explanations (LIME)

Ribeiro, Singh, and Guestrin 2016


A déjà vu?
Same basic
question, slightly
different context?

Design and Analysis of Computer Experiments (DACE)

Situation: Real life experiments replaced by computer simulations



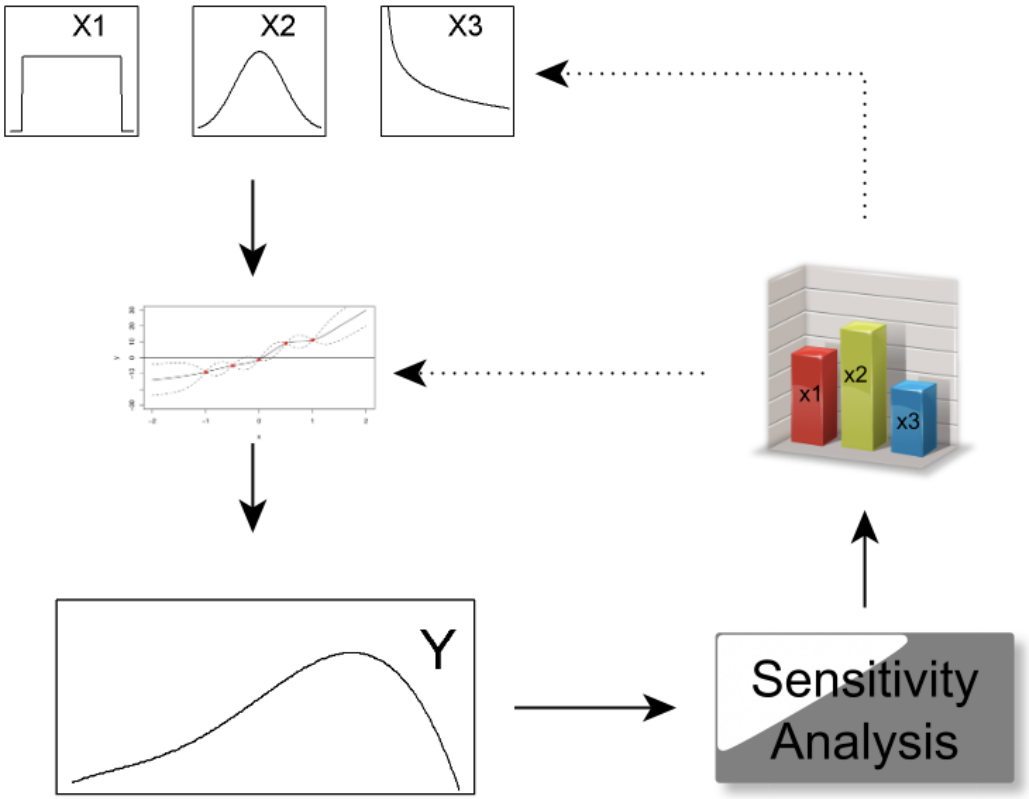
Metamodel: Complex and cost/time efficient simulation model are modeled by e.g. Gaussian process models (Kriging)

Analyze metamodel $f(X)$ instead of simulation model:

- Identify input variables which strongly affect the variation in the output

Global Sensitivity Analysis (GSA)

Sensitivity analysis is the study of how the variation in the output of a model can be apportioned to different sources of variation



Sobol Indices play vital role in variance-based sensitivity analysis

Saltelli, Chan, and Scott 2000

Overview

- Sobol Indices and FanovaGraph
- Shapley Values and FaithShapGraph
- California Housing Example

Sobol Indices

Functional ANOVA decomposition of a function

$f : \Delta \to \mathbb{R}$, $f \in L_2(\Delta, \mathbb{R})$ with $\Delta = \Delta_1 \times \ldots, \times \Delta_d$ continuous function.

$$f(X) = f_0 + \sum_{i=1}^{d} f_i(X_i) + \sum_{i<j} f_{i,j}(X_i, X_j) + \cdots + f_{1,\ldots,d}(X_1, \ldots, X_d)$$

Properties:

$$\forall J : E(f_J(X_J)) = 0,$$
$$\forall J' \neq J : \; E(f_J(X_J) f_{J'}(X_{J'})) = 0$$

Taking the variance yields the variance decomposition:

$$D := \mathrm{Var}(f(\mathbf{x})) = \mathrm{Var}(f_0) + \sum_{i=1}^{d} \mathrm{Var}(f_i(X_i)) + \sum_{1 \leq i < j \leq d} \mathrm{Var}(f_{ij}(X_i, X_j)) + \cdots + \mathrm{Var}(f_{1,\ldots,d}(X_1, \ldots, X_d))$$

Sobol 1993

Sobol Indices

Sobol indices quantify the contribution of input variables X_i to the variance of the model output Y

- **First-Order Index:**

$$D_i := \mathrm{Var}(f_i(X_i)), \quad S_i := \frac{D_i}{D},$$

for variable X_i, where $D = \mathrm{Var}(f(X))$

- **Total interaction index (TII)**:
Joint contribution to the variance of Y by two variables X_i and X_j

$$D_{ij} := \mathrm{Var}\left(\sum_{I \supseteq \{i,j\}} \mu_I(X_I)\right) = \sum_{I \supseteq \{i,j\}} D_I$$

Fruth and Kuhnt 2012

Estimation

Estimation method *fixLO* by Monte Carlo integration

$$\widehat{\mathfrak{D}}_{ij} = \frac{1}{n} \cdot \frac{1}{4} \sum_{m=1}^{n} \Big[f(x_i^m, x_j^m, x_{-\{i,j\}}^m) - f(x_i^m, z_j^m, x_{-\{i,j\}}^m) - f(z_i^m, x_j^m, x_{-\{i,j\}}^m) + f(z_i^m, z_j^m, x_{-\{i,j\}}^m) \Big]^2$$

with $(x_i^m, x_j^m, x_{-\{i,j\}}^m)'$, $(x_i^m, z_j^m, x_{-\{i,j\}}^m)'$, $(z_i^m, x_j^m, x_{-\{i,j\}}^m)'$, and $(z_i^m, z_j^m, x_{-\{i,j\}}^m)'$

four MC-samples of $\mathbf{X}$ for $m = 1, \ldots, n$.

Liu and Owen 2006

Ishigami Function Example

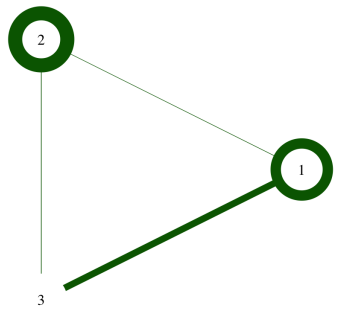
The Ishigami function is a popular benchmark for sensitivity analysis.

$$f(x_1, x_2, x_3) = \sin(x_1) + 7\,\sin^2(x_2) + 0.1\,x_3^4\,\sin(x_1)$$

with x_i uniformly distributed on $[-\pi, \pi]$

Estimation for Ishigami function(Latin Hypercube Design, n=100)

$\{i\}$	$\widehat{\mathfrak{D}}_i$	$\mathfrak{D}_i$
$\{1\}$	4.116	4.346
$\{2\}$	6.384	6.125
$\{3\}$	0.041	0
$\{i,j\}$	$\widehat{\mathfrak{D}}_{ij}$	$\mathfrak{D}_{ij}$
$\{1,3\}$	3.133	3.375
$\{1,2\}$	0	0
$\{2,3\}$	0	0



Estimated FANOVA graph

Thresholding necessary, for example 0.1

Ishigami and Homma 1990; Fruth and Kuhnt 2012

Overview

- Sobol Indices and FanovaGraph
- Shapley Values and FaithShapGraph
- California Housing Example

Shapley Values

Shapley values, originating from cooperative game theory, provide a fair way to attribute the contribution of each feature to a model's prediction.

$$\phi_i(f, x) = \sum_{S \subseteq V \setminus \{i\}} \frac{|S|!\,(|V| - |S| - 1)!}{|V|!} \Big(f_{S \cup \{i\}}(x) - f_S(x) \Big)$$

- V: set of all features, $V = 1, 2, ..., d$
- S: any subset of features excluding i
- $f_S(x)$: model prediction using features in S

Shapley 1953

Shapley Values: Local vs. Global

Local explanation

- Single prediction x
- Coalition value:

$$v(S) = f(x_S, x^*_{\bar{S}})$$

- Missing features $x_{\bar{S}}$ imputed

Global explanation

- Aggregate over dataset $\{x^{(m)}\}_{m=1}^{M}$
- Average contributions:

$$\Phi_j = \frac{1}{M} \sum_{m=1}^{M} \phi_j(x^{(m)})$$

Global Shapley values comparable to first-order Sobol Indices, but what about interactions?

Shap Value Interaction Index

Recently: Extensions of Shapley framework to capture feature interactions, quantifying how groups of features jointly affect the model prediction

- The Shapley Taylor Interaction Index
  (Sundararajan, Dhamdhere, and Agarwal 2020)
- Faith-shap: The faithful shapley interaction index
  (Tsai, Yeh, and Ravikumar 2023)
- SHAP-IQ: Unified approximation of any-order shapley interactions
  (Fumagalli et al. 2023)

FaithShap for Interaction Indices

- **Joint effect beyond individual contributions:** FaithShap assigns a contribution $\Phi_k(S)$ to each coalition $S \subseteq J$ of features, up to an *explanation order* k
- **Faithfulness objective:**

$$L(\nu, \Phi_k) = \sum_{T \subseteq J} \mu(t) \left(\nu(T) - \sum_{S \subseteq T, |S| \leq k} \Phi_k(S) \right)^2$$

with

$$\mu(t) := \begin{cases} \mu_\infty, & \text{if } t \in \{0, d\}, \\ \frac{1}{\binom{n-2}{t-1}}, & \text{else} \end{cases}$$

is minimized to optimize Shapley-weighted faithfulness

- FaithShap guarantees that the sum of individual and interaction contributions equals the total prediction difference

Tsai, Yeh, and Ravikumar 2023; Fumagalli et al. 2023

FaithShap Graph

Idea: Introduce a **FaithShap Graph** comparable to FanovaGraph to visualise

- **SHAP Values** quantifying the contributions of individual features
- **Feature Interactions** illustrating how combinations of features jointly influence model predictions

Again the visualization encodes:

- **Node Sizes** corresponding to the magnitude of each feature's main effect.
- **Edge Thicknesses** representing the strength of interactions between features

FaithShap Graph for Ishigami Function

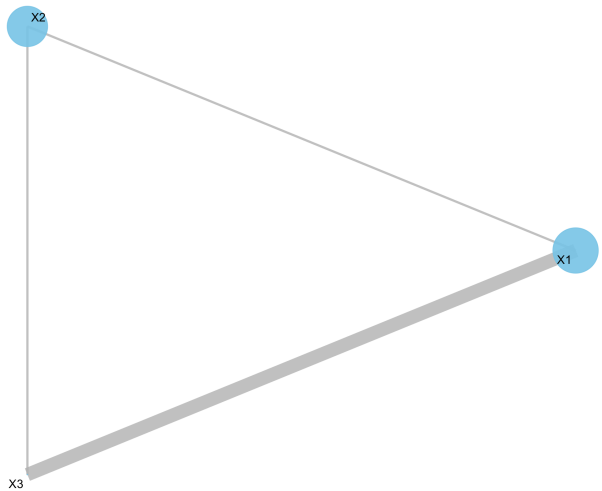
The Ishigami function is defined as:

$$f(x_1, x_2, x_3) = \sin(x_1) + 7\,\sin^2(x_2) + 0.1\,x_3^4\,\sin(x_1)$$

with x_i uniformly distributed on $[-\pi, \pi]$.
Estimation for Ishigami function(Latin Hypercube Design, n=100)

$\{i\}$	Φ_i
Φ_1	1.88
Φ_2	2.23
Φ_3	0.24
$\{i,j\}$	Φ_{ij}
$\Phi_{1,3}$	1.34
$\Phi_{1,2}$	0.15
$\Phi_{2,3}$	0.15

Python package `shapiq` used to calculate values



Estimated FaithShap graph for the Ishigami function.

Overview

- Sobol Indices and FanovaGraph
- Shapley Values and FaithShapGraph
- California Housing Example

Real World Example: California Housing

- Data from 1990 U.S. Census on StatLib repository.
- 20,640 observations, each representing a California census block group
- Block groups are the smallest geographic units published by the Census Bureau (typically 600–3,000 people)
- Target variable: Median house value (in $100,000 units)

Pedregosa et al. 2011

Real World Example: California Housing

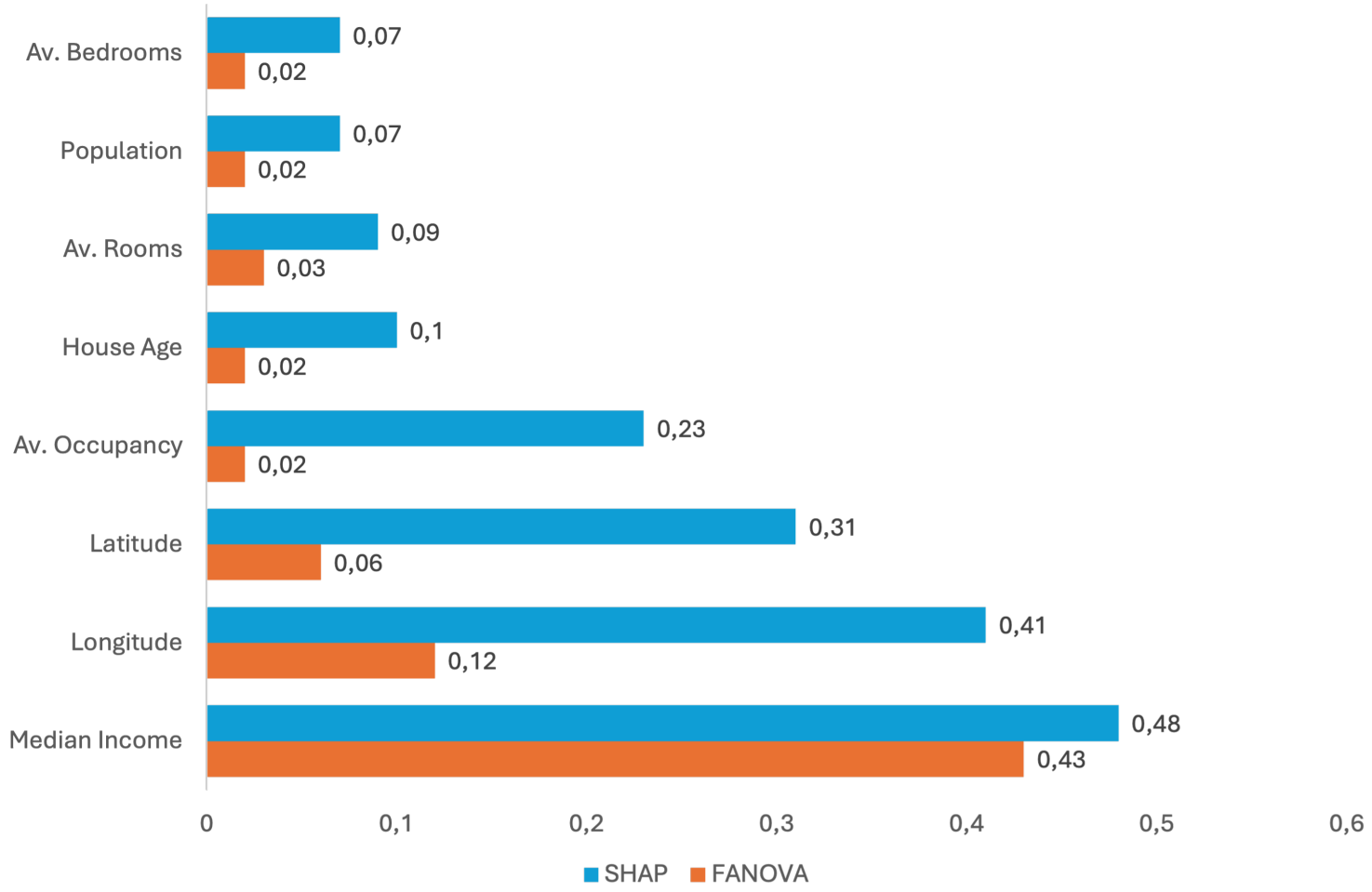
MedInc	Median income of households in the block group in $10.000
HouseAge	Median age of houses in the block group (years)
AveRooms	Average number of rooms per household (count)
AveBedrms	Average number of bedrooms per household (count)
Population	Total population of the block group (persons)
AveOccup	Average number of people per household (persons)
Latitude	Geographical latitude of the block group (decimal degrees)
Longitude	Geographical longitude of the block group (decimal degrees)

Pedregosa et al. 2011

Real World Example: California Housing

$f(X)$ given by **Random Forest Regressor** predicting median house prices

Comparing First Order Sobol Indices and Shapley Values

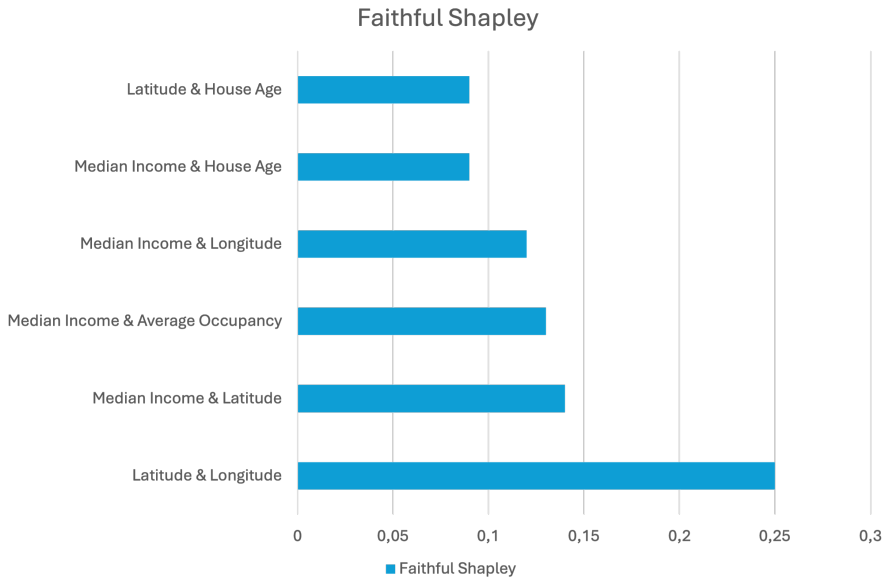


Real World Example: California Housing

$f(X)$ given by **Random Forest Regressor** predicting median house prices

Comparing Total Interaction Indices and Faithful Shapley Interactions

Total Interaction Indices



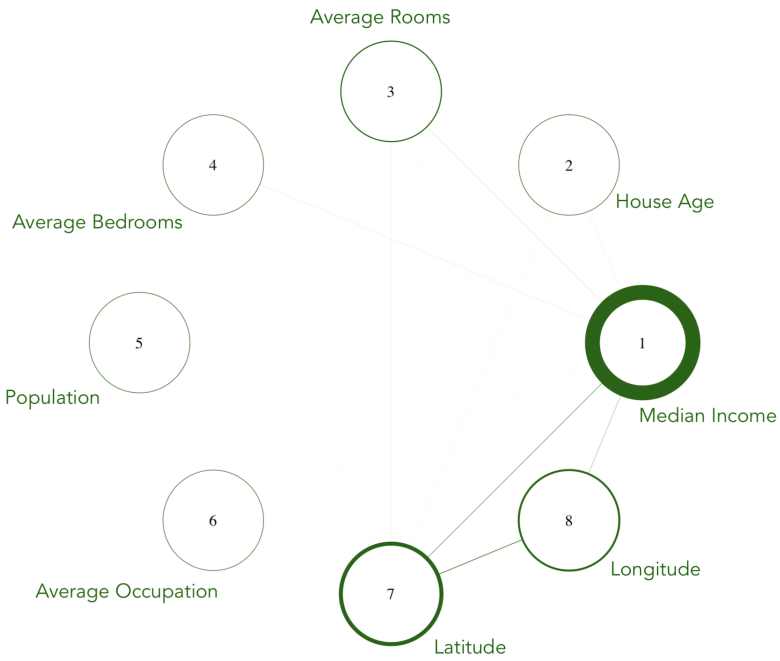
Faithful Shapley Interactions

Both highlight **Latitude & Longitude** as the strongest interaction.
Median Income & Latitude appear in both as well.

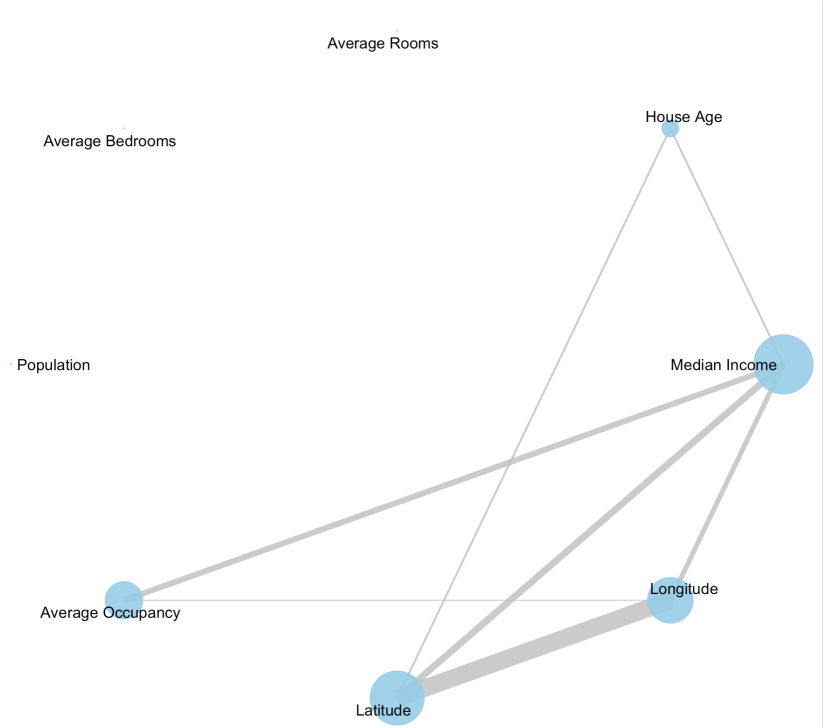
Real World Example: California Housing

Comparing Total Interaction Indices and Faithful Shapley Interactions

$f(X)$ given by **Random Forest Regressor** predicting median house prices



FANOVA graph



FaithShap graph

Summary

Why Use Both Methods in XAI?

- Explainability ensures trust, fairness, and transparency in AI
- Sobol (sensitivity analysis) and Shapley (game theory) offer complementary views on feature importance
- Combining both enhances ML model interpretability

Limitations & Future Work

- Both methods have computational costs: SHAP is expensive for large datasets
- Sobol indices assume independent inputs, which may not hold in real-world datasets
- Explore and compare theoretical properties

References I

Fruth, J. and S. Kuhnt (2012). "Sensitivity Analysis and FANOVA Graphs for Computer Experiments". In: *46TH SCIENTIFIC MEETING OF THE ITALIAN STATISTICAL SOCIETY*.

Fumagalli, F. et al. (2023). "SHAP-IQ: Unified approximation of any-order shapley interactions". In: *Advances in Neural Information Processing Systems* 36, pp. 11515–11551.

Ishigami, T. and T. Homma (1990). "An Importance Quantification Method in Global Sensitivity Analysis". In: *Proceedings of the 2nd International Symposium on Uncertainty Modeling and Analysis*, pp. 398–403.

Liu, R. and A. B. Owen (2006). "Estimating mean dimensionality of analysis of variance decompositions". In: *Journal of the American Statistical Association* 101.474, pp. 712–721.

Muehlenstaedt, T. et al. (2012). "Data-driven Kriging models based on FANOVA-decomposition". In: *Stat Comput* 22, pp. 723–738.

Muschalik, M. et al. (2024). "shapiq: Shapley Interactions for Machine Learning". In: *The Thirty-eight Conference on Neural Information Processing Systems (NeurIPS 2024)*.

Owen, A. B. (2014). "Sobol' indices and Shapley value". In: *SIAM/ASA Journal on Uncertainty Quantification* 2, pp. 245–251.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

References II


Ribeiro, M., S. Singh, and C. Guestrin (2016). ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 1135–1144.


Saltelli, A., K. Chan, and E. M. Scott (2000). *Sensitivity analysis*. Wiley.


Shapley, L. S. (1953). "A value for n-person games". In: *Contributions to the Theory of Games* 28, p. 307.


Sobol, I. M. (1993). "Sensitivity estimates for nonlinear mathematical models". In: *Mathematical Modeling and Computational Experiment* 1, pp. 407–414.


Sundararajan, M., K. Dhamdhere, and A. Agarwal (2020). "The Shapley Taylor Interaction Index". In: *International Conference on Machine Learning*, pp. 9259–9268.


Tsai, C., C. Yeh, and P. Ravikumar (Jan. 2023). "Faith-shap: the faithful shapley interaction index". In: *J. Mach. Learn. Res.* 24.1. ISSN: 1532-4435.