Shedding Light on Black Box Machine Learning Algorithms

Development of an Axiomatic Explanation Consistency Framework to Assess the Quality of Methods that Explain Individual Predictions

Milo R. Honegger

Reviewer:	Prof. Dr. rer. pol. Christof Weinhardt
Second reviewer:	Prof. Dr. Alexander Maedche
Advisor:	Rico Knapper
Second advisor:	Dr. Sebastian Blanc

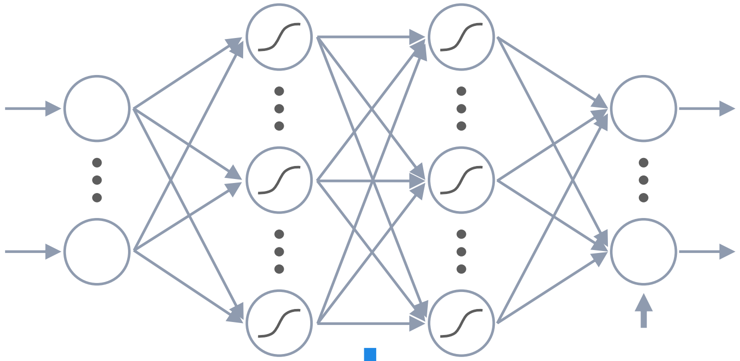
27.06.2018

The Classical Machine Learning Task

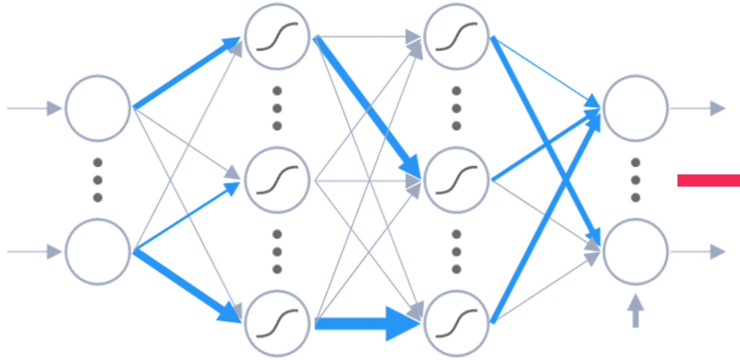
House Sales Price Dataset

House	Size (m^2)	Location (lat/long)	Year built	Condition	⋯	Price ($)
1	220	47,23 / -122,10	1995	Good	⋯	430k
2	150	47,58 / -122,23	1987	Ok	⋯	250k
3	340	47,92 / -122,55	2009	New	⋯	740k
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Prediction Model



Trained Prediction Model



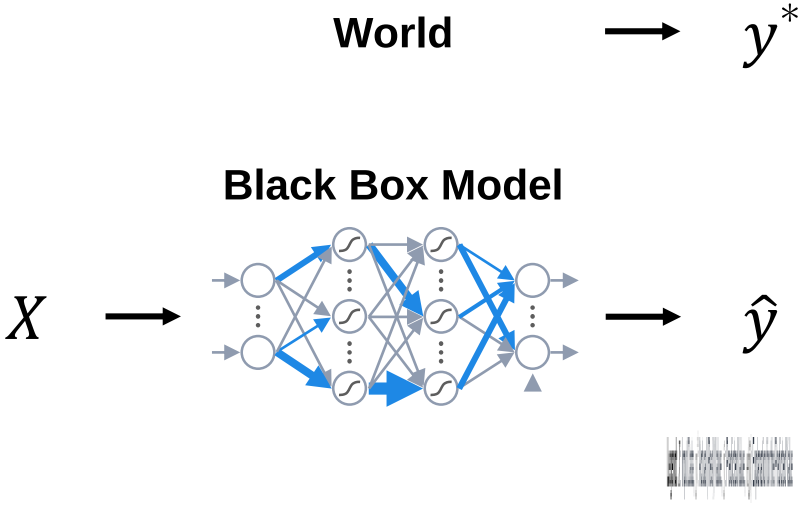
New Housing Dataset

House	Size (m^2)	Location (lat/long)	Year built	Condition	⋯	Price ($)
1	245	47,85 / -122,92	1992	Ok	⋯	??
2	150	47,28 / -122,48	1997	Good	⋯	??
3	340	47,95 / -122,73	2011	New	⋯	??
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Predicted House Price

$ 615.484

Black Box Machine Learning Models and Explanation Methods

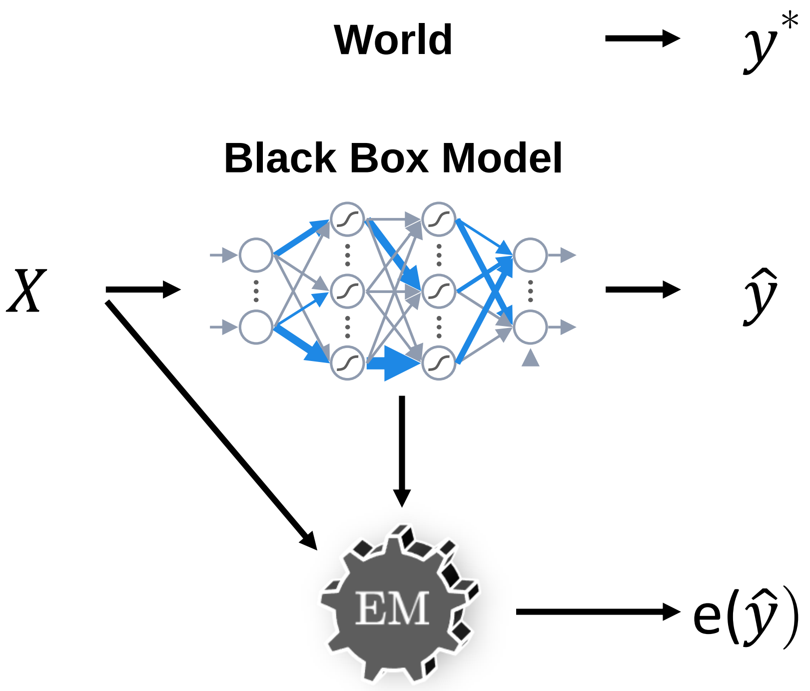


Problem: Loss of Interpretability

Model Complexity ↗

[Accuracy ↗ *] ?*

Interpretability ↘



Solution: Explanation Methods

Model Complexity ↗

[Accuracy ↗ *] ?*

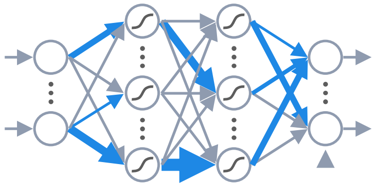
Interpretability ↗

Legend: X: Input Data, y^*: Actual (Real) Value, $\hat{y}$: Predicted Value, e($\hat{y}$): Explanation for the Predicted Value

Explanation: “The Answer to a Why-Question” (Miller, 2017) (1/2)

House in King County

Black Box Model



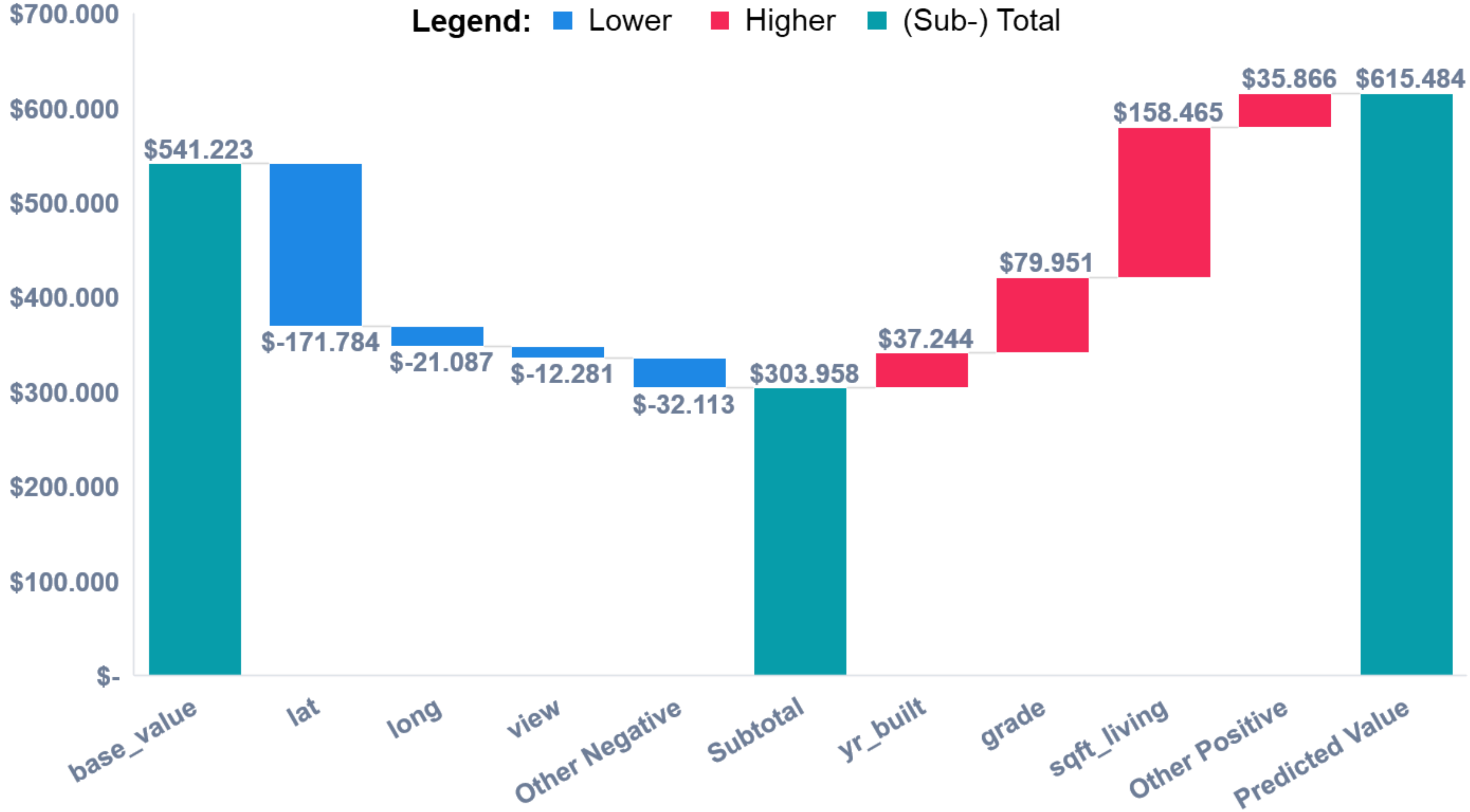
Predicted House Price

$ 615.484

Why?

Explanation

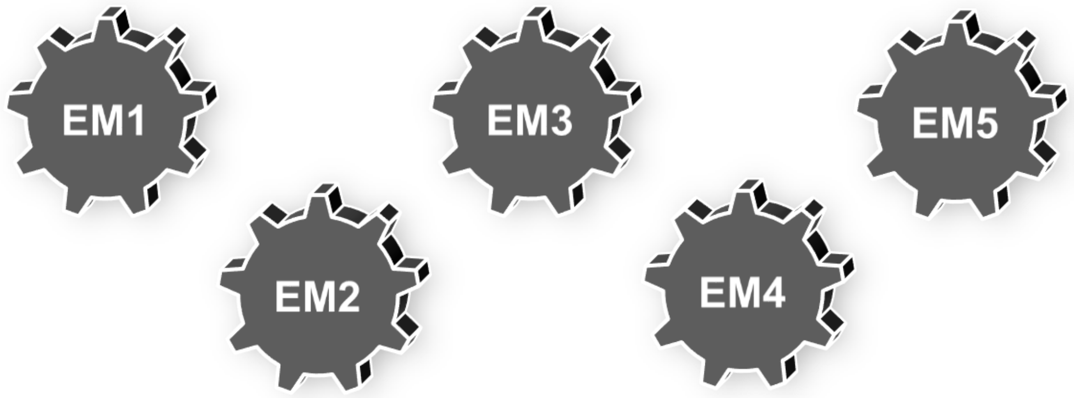
(Shapley Additive Explanations)



[ænə'sɪʒn]

Research Gap & Research Question

Current Research Focus & Gap



Current Research Focus:
Developing new explanation methods

Research Gap:
Method to compare and assess the quality, strengths and weaknesses of different EMs

Research Question

"Can we develop an **axiomatic framework** to assess the **quality of explanation methods**, used to explain individual predictions made by **black box machine learning** models?"

[ænə'sɪʒn]

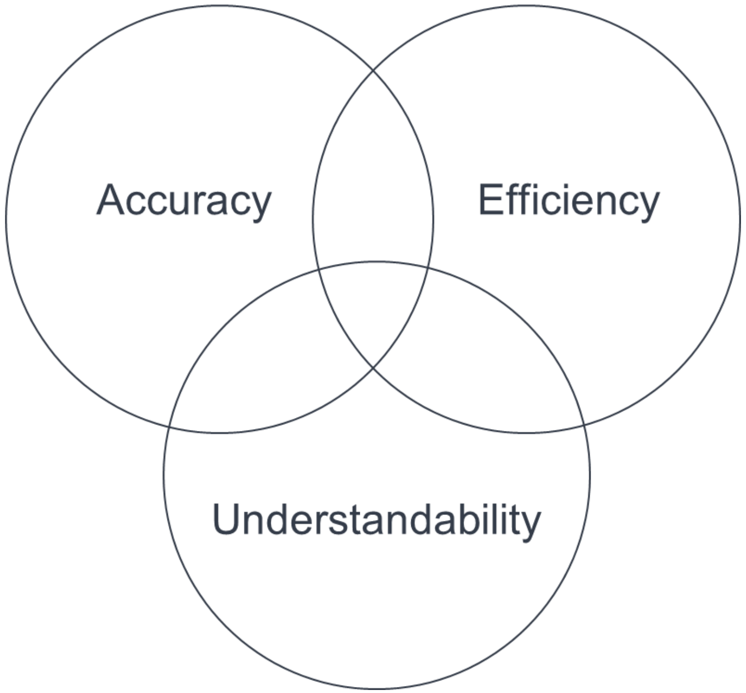
Agenda

Agenda

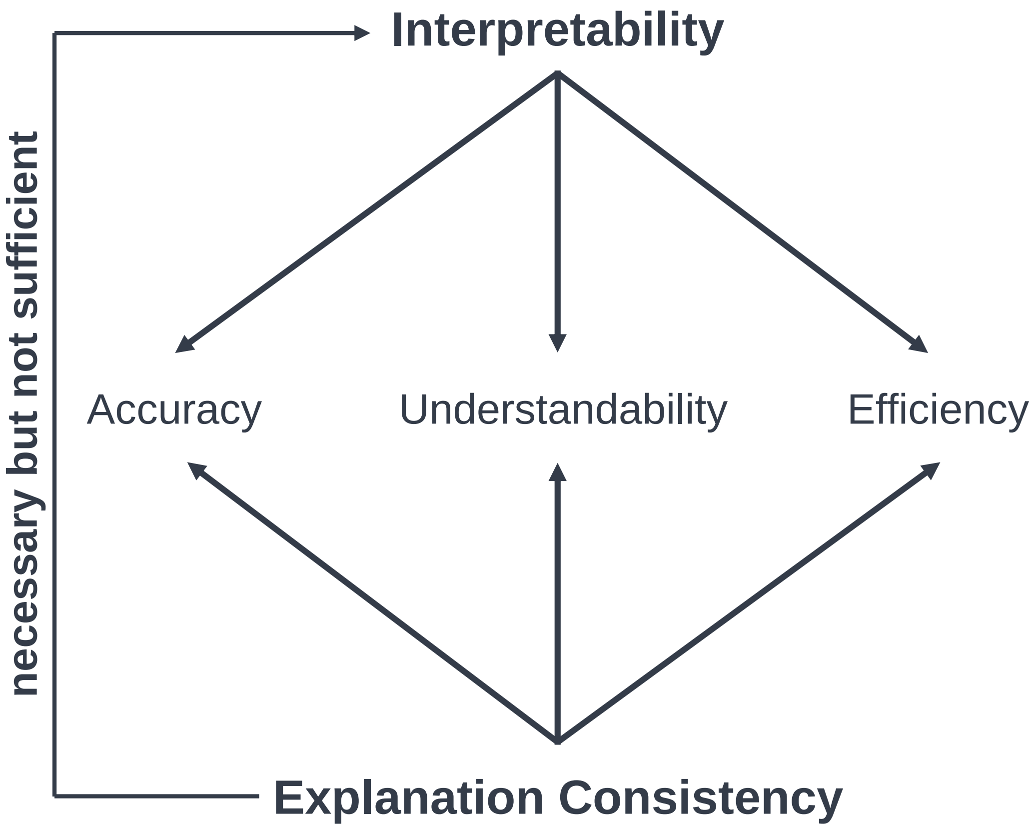
Interpretability and Explanation Consistency

Interpretability as a Quality Measure

The Goals of Interpretability
(based on Rüping et al., 2006)



Explanation Consistency Framework



[ænə'sɪʒn]

Axiomatic Explanation Consistency (Regression Case)

Axioms

1. **Identity:** Identical objects must have identical explanations:

$$d(\vec{x}_a, \vec{x}_b) = 0 \Longrightarrow d(\vec{\varepsilon}_a, \vec{\varepsilon}_b) = 0,$$
$$\forall\, a, b$$

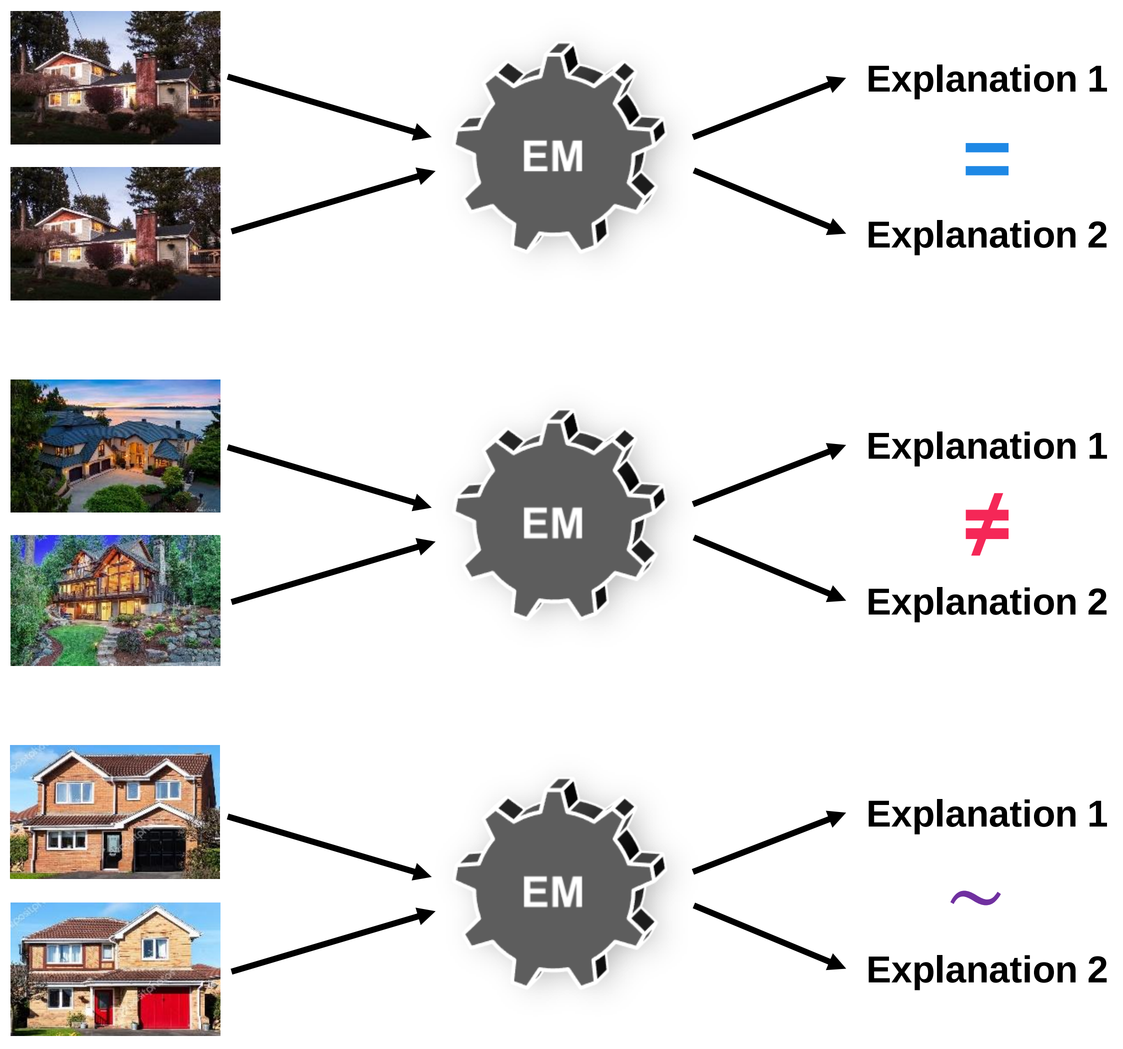
2. **Separability:** Non-identical objects can not have identical explanations:

$$d(\vec{x}_a, \vec{x}_b) \neq 0 \Longrightarrow d(\vec{\varepsilon}_a, \vec{\varepsilon}_b) > 0,$$
$$\forall\, a, b$$

3. **Stability:** Similar objects must have similar explanations:

$$\rho\left(D_{Z\,j}, D_{E\,j}\right) = \rho_j > 0,$$
$$\forall\, j \in |Z|, \qquad \rho_j \subset P$$

Illustration



Agenda

Experiments: Seattle House Prices

Dataset

- Homes in King County (Seattle)
- >21.000 house sale prices (2014 / 2015)
- 19 features for each house
- Mean Price $540.607

Prediction Model

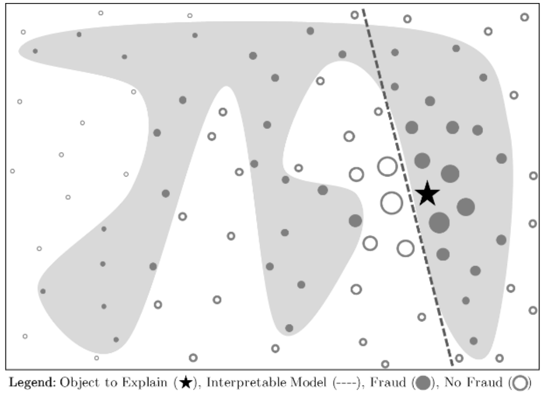
- Extreme Gradient Boosting (XGB)
- Tree ensemble
- 100 boosting rounds (estimators)
- Overfitting control

Model Performance

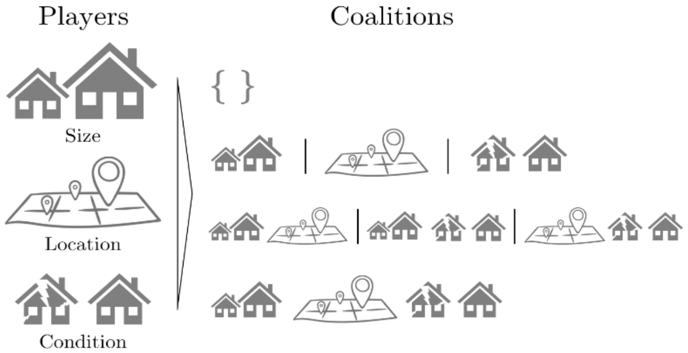
- R^2: 0.89
- RMSE: $120.478
- 22,29% deviation from mean price
- Outliers influence on performance measures

Explanation Methods

LIME (Ribeiro et al., 2016)
(Local Interpretable Model-Agnostic Explanations)



SHAP (Lundberg et al., 2017)
(Shapley Additive Explanations)



Experiments: Evaluation of the Explanation Consistency of LIME and SHAP

Explanation Method	Axiom	# violated	# satisfied	% satisfied
LIME	**1. Identity**	5.355	0	**0%**
	2. Separability	**134**	28.670.536	99,9995%
	3. Stability	**4**	5.351	99,9252%
SHAP	**1. Identity**	0	5.355	**100%**
	2. Separability	**28**	28.670.642	99,9999%
	3. Stability	**0**	5.355	100%

Stability Axiom: in-depth Analysis

Spearman's Rho	LIME	SHAP
Minimum	-0,0086	0,1685
Maximum	0,8001	0,8934
Mean	0,4902	0,7020
Median	0,5037	0,7150

Experiments: Strengths, Weaknesses and Application Domains

LIME: Strengths
+ Fast + Returns a prediction model + Tabular, text and image data

SHAP: Strengths
+ Solid theoretical foundation + High explanation consistency + Lightning fast for tree ensembles

LIME: Weaknesses
- Randomness (identity axiom) - Specific knowledge required - No solid theory

SHAP: Weaknesses
- Exponential complexity for non-tree ensembles $O(2^k)$ - Misinterpretation - Does not return a prediction model

LIME: Best Application Domains
▪ Most real world problems where an approximate solution suffices ▪ Datasets with a big amount of features

SHAP: Best Application Domains
▪ Situations that demand explainability by law (GDPR) ▪ Debugging (model / data bias)

Agenda

Discussion of Limitations & Outlook

Limitations: Approach

- Test other EMs
- Validate with more datasets
- People validation

Future Directions for Research

- Validation with more EMs, datasets and people
- Further axioms
- Variance-based approach

Limitations: Explanation Consistency Framework

- Stricter sub-axioms?
- Computational complexity
- Aggregated distances = loss of information?

Outlook and Next Steps

- Research Paper
- Explanation Demonstrator Dash

Agenda

[ænəˈsɪʒn]

Conclusion

Research Question

“Can we develop an **axiomatic framework** to assess the **quality of explanation methods**, used to explain individual predictions made by **black box machine learning** models?”

Bottom Line

- **Yes**!
- The explanation consistency framework represents **groundbreaking** research
- It is a **starting point** to motivate **further research**
- Useful and feasible to **measure** and **compare explanation quality**
- **Work in Progress**

[ænəˈsɪʒn]

Thank You very much for your Attention!

Main Literature Sources (not exhaustive)

- **Molnar, C., et al. (2018)**. Interpretable machine learning: A guide for making black box models explainable.
- **Rüping, S., et al. (2006)**. Learning interpretable models.
- **Friedman, J., Hastie, T., & Tibshirani, R. (2009)**. The elements of statistical learning 2nd edition. New York: Springer.
- **Miller, T. (2017)**. Explanation in artificial intelligence: Insights from the social sciences. arXiv preprint arXiv:1706.07269 .
- **Sundararajan, M., Taly, A., & Yan, Q. (2017)**. Axiomatic attribution for deep networks. arXiv preprint arXiv:1703.01365.
- **Doshi-Velez, F., & Kim, B. (2017)**. Towards a rigorous science of interpretable machine learning.
- **Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., . . . Wood, A. (2017)**. Accountability of ai under the law: The role of explanation. arXiv preprint arXiv:1711.01134 .
- **Lundberg, S. M., & Lee, S.-I. (2017a)**. Consistent feature attribution for tree ensembles. arXiv preprint arXiv:1706.06060.
- **Lundberg, S. M., & Lee, S.-I. (2017b)**. A unified approach to interpreting model predictions. In Advances in neural information processing systems (pp. 4768{4777).
- **Ribeiro, M. T., Singh, S., & Guestrin, C. (2016)**. Why should i trust you?: Explaining the predictions of any classier. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 1135{1144).