

Confident Interpretations of Black Box classifiers

Nedeljko Radulović
Albert Bifet Fabian Suchanek

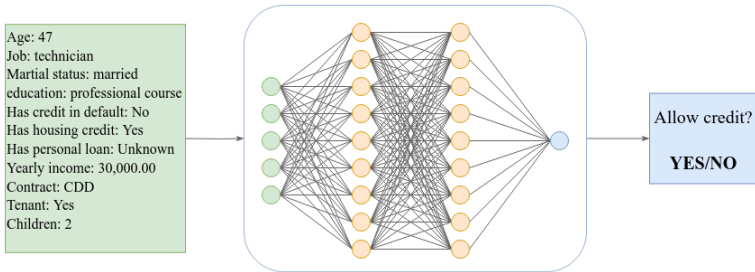
radulovic.nedeljko@telecom-paris.fr
Télécom Paris, Institut Polytechnique de Paris



May 20, 2021

Use case scenario

- Use Neural Network to answer a loan request



Introduction

Related Work

STACI:
Surrogate
Trees for A
posteriori
Confident
Interpretations

Experimental
results

Summary

Use case scenario

Introduction

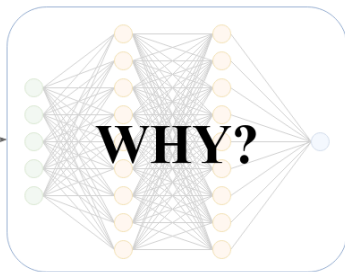
Related Work

STACI:
Surrogate
Trees for A
posteriori
Confident
Interpretations

Experimental
results

Summary

Age: 47
Job: technician
Marital status: married
education: professional course
Has credit in default: No
Has housing credit: Yes
Has personal loan: Unknown
Yearly income: 30,000.00
Contract: CDD
Tenant: Yes
Children: 2



Allow credit?

NO

Explainable Artificial Intelligence

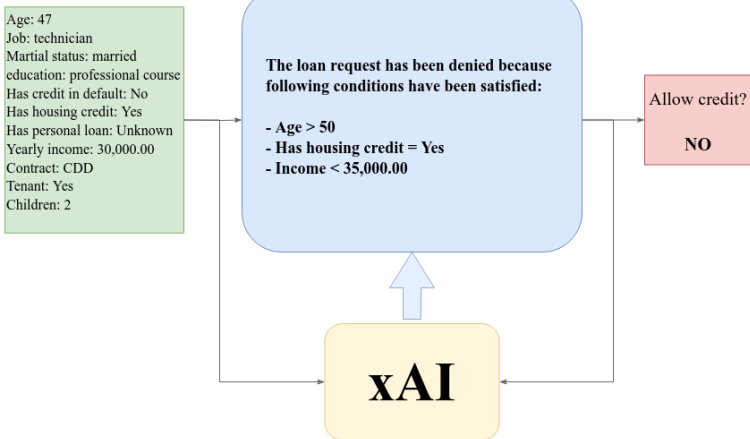
Introduction

Related Work

STACI:
Surrogate
Trees for A
posteriori
Confident
Interpretations

Experimental
results

Summary



Related work

Introduction

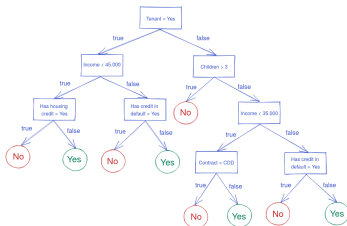
Related Work

STACI:
Surrogate
Trees for A
posteriori
Confident
Interpretations

Experimental
results

Summary

Building already interpretable models : Decision trees, Rule-based models and linear models



Decision Tree

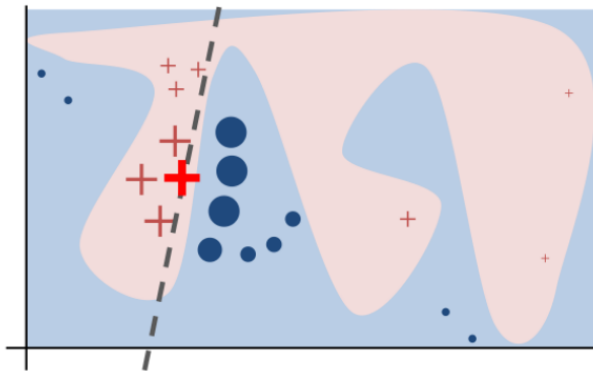
- If Age > 48, then **No**,
- else if Has housing credit = Yes, then **No**,
- else if Children > 3, then **No**,
- else if Has credit in default = Yes, then **No**,
- else if Income < 40.000, then **No**,
- else if Contract = CDD, then **No**,
- else **Yes**.

Rule based model

Related work

Post-hoc interpretability: Building surrogate interpretable models

- Local models: LIME [1], Anchors [2], SHAP [3]
- Global models: TREPAN [4], DTEExtract [5]



Introduction

Related Work

STACI:
Surrogate
Trees for A
posteriori
Confident
Interpretations

Experimental
results

Summary

Approximation using interpretable model

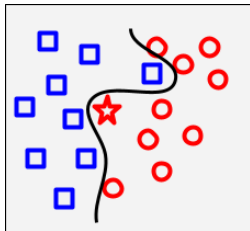
Introduction

Related Work

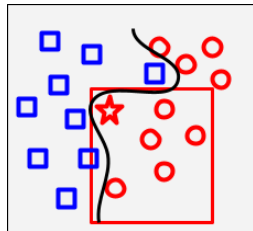
STACI:
Surrogate
Trees for A
posteriori
Confident
Interpretations

Experimental
results

Summary

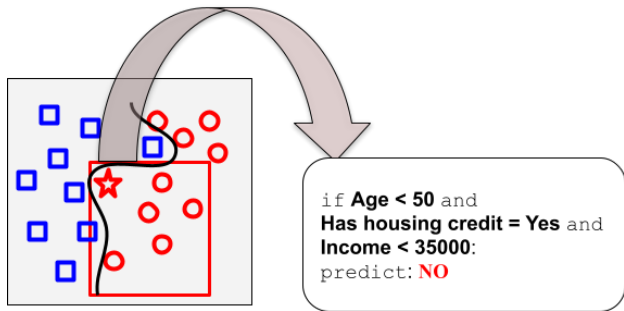


Black box model decision
boundary



Interpretable model
approximation

Interpretation

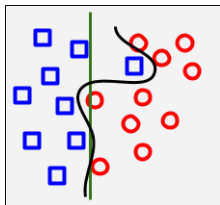


Interpretation provided by interpretable model

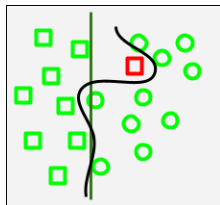
4 Criteria

First two criteria are common:

- **Complexity** - Length of the interpretation
- **Fidelity** - Interpretable model is faithful to the black box model



Complexity

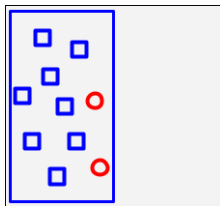


Fidelity

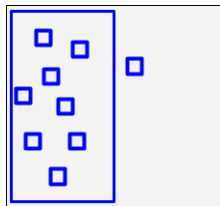
4 Criteria

We introduce **two new** criteria:

- **Confidence** - Interpretation applies on data points of the same class
- **Generality** - Interpretation applies on multiple data points



Confidence



Generality

The main idea

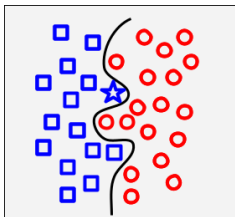
Introduction

Related Work

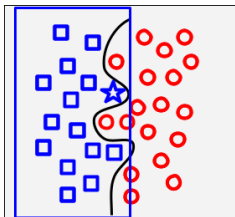
STACI:
Surrogate
Trees for A
posteriori
Confident
Interpretations

Experimental
results

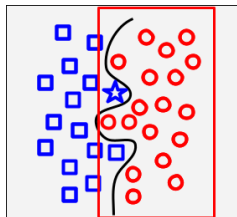
Summary



Original black box
model



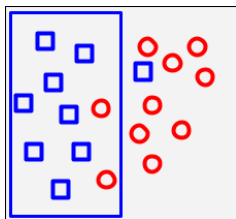
Interpretable model
for left class



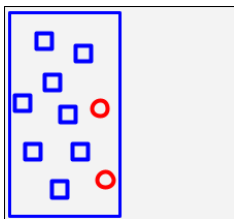
Interpretable model
for right class

Training

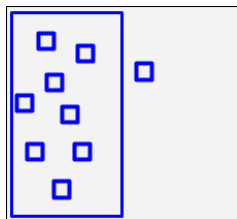
- Decision tree as interpretable model
- **Complexity** - Define the maximal length of the interpretation
- **Fidelity** - Label training data using the black box model
- Use *F1* measure as a metric for deciding a split:
 - Confidence \leftrightarrow Precision
 - Generality \leftrightarrow Recall



Interpretable model



Confidence



Generality

Introduction

Related Work

STACI:
Surrogate
Trees for A
posteriori
Confident
Interpretations

Experimental
results

Summary

Fidelity

Table: Fidelity (%) with NN as black box model

Dataset	DTE	SBRL	LIME	CART	STACI'	STACI
Heart	87.34	85.88	84.84	80.97	79.68	84.84
Breast	94.93	91.57	87.28	89.65	91.05	93.16
Diabetes	80.58	83.38	71.49	75.19	76.23	84.55
Voting	95.91	94.55	95.34	95.34	94.55	95.00
Sick	97.88	97.25	75.36	96.66	97.79	98.46
Hypo.	96.39	97.88	94.32	98.99	98.45	99.31
Adult	92.35	93.88	87.56	73.53	98.23	99.58
Wine	91.11	N/A	52.78	66.67	86.67	97.78
Derma.	94.86	N/A	82.70	80.28	95.28	96.11
Vehicle	74.47	N/A	54.71	69.06	68.24	86.35

Complexity

Table: Average Complexity

Dataset	Black	DTE	SBRL	LIME	CART	STACI
Heart	NN	3.15	3.90	3	3	2.89
	RF	3.11	2.29	4	4	3.28
Breast	NN	2.88	4.20	3	3	1.9
	RF	3.18	6.16	4	4	2.88
Diabetes	NN	2.89	5.78	3	3	1.49
	RF	2.75	7.21	4	4	1.85
Voting	NN	3.11	1.57	3	3	1.58
	RF	3.00	1.63	3	3	1.69
Sick	NN	2.40	3.64	3	3	1.40
	RF	2.25	3.77	3	3	2.07
Hypo.	NN	2.58	4.50	3	3	1.20
	RF	2.16	4.78	3	3	1.09
Adult	NN	3.25	8.49	4	4	1.87
	RF	2.75	7.22	4	4	1.83
Wine	NN	3.95	N/A	3	3	2.42
	RF	4.29	N/A	4	4	2.93
Derma.	NN	4.91	N/A	3	3	2.24
	RF	4.85	N/A	4	4	2.36
Vehicle	NN	3.99	N/A	3	3	2.68
	RF	4.50	N/A	4	4	2.91

Introduction

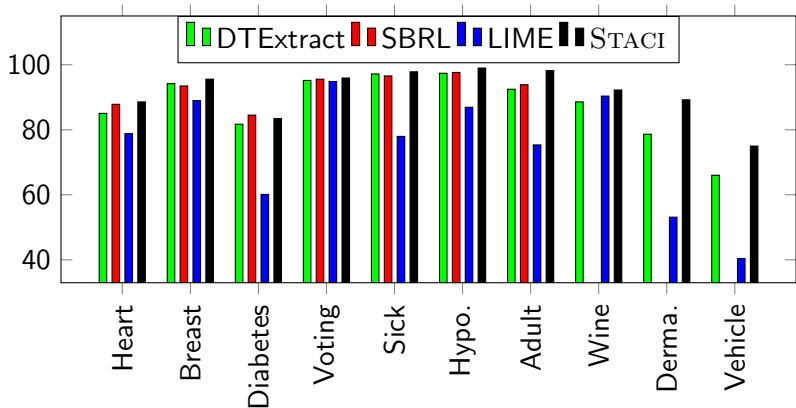
Related Work

STACI:
Surrogate
Trees for A
posteriori
Confident
Interpretations

Experimental
results

Summary

Confidence



Introduction

Related Work

STACI:
Surrogate
Trees for A
posteriori
Confident
Interpretations

Experimental
results

Summary

Generality

Table: Generality comparison

Dataset	Black	DTE	STACI
Heart	NN	59.21	76.63
	RF	58.83	68.35
Breast	NN	80.31	92.59
	RF	84.82	88.67
Diabetes	NN	66.92	74.47
	RF	64.23	71.51
Voting	NN	73.37	95.01
	RF	82.14	95.15
Sick	NN	94.70	94.18
	RF	93.39	94.43
Hypo.	NN	89.62	97.08
	RF	96.79	96.62
Adult	NN	92.06	95.53
	RF	92.25	73.84
Wine	NN	77.03	86.67
	RF	79.51	85.12
Derma.	NN	91.74	91.33
	RF	91.54	91.54
Vehicle	NN	53.98	68.70
	RF	46.16	55.54

Interpretation example

The datapoint

Pregnancies	5
Glucose	166
Blood pressure	72
Skin thickness	19
Insulin	175
BMI	25.8
Diabetes pedigree	0.59
Age	51

is classified as diabetic. It has these characteristics:

Glucose > 154, Insulin > 145, Age > 30

There are 37 other data points with these characteristics, and 94.59% of them are also classified as diabetic.

Introduction

Related Work

STACI:

Surrogate

Trees for A

posteriori

Confident

Interpretations

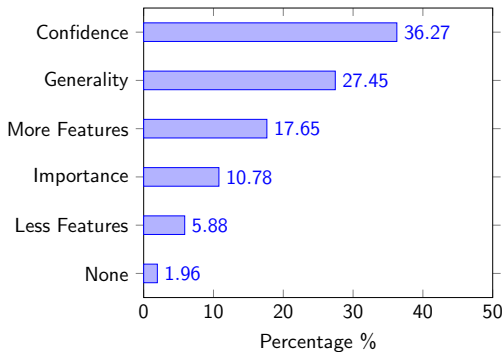
Experimental

results

Summary

User study

System	Confidence(%)	Generality(%)	Length	Average Rating
DTEExtract	76.61	74.68	1.16	3.12
LIME	59.18	0.41 ¹	1.86	1.93
STACI	85.23	42.42	2.52	3.91



¹Local Model

STACI: Surrogate Trees for A posteriori Confident Interpretations

Introduction

Related Work

STACI:
Surrogate
Trees for A
posteriori
Confident
Interpretations

Experimental
results





Summary

Summary:

- Train one decision tree per class using $F1$ as a metric for a split
- Provide: confident, general and simple interpretations

Future works:

- Remove the need for the user defined maximal length

-  M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should i trust you? – explaining the predictions of any classifier,” in *SIGKDD*, 2016.
-  —, “Anchors: High-precision model-agnostic explanations,” in *AAAI*, 2018.
-  S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in neural information processing systems*, 2017.
-  M. Craven and J. W. Shavlik, “Extracting tree-structured representations of trained networks,” in *Advances in neural information processing systems*, 1996.

Introduction

Related Work

STACI:
Surrogate
Trees for A
posteriori
Confident
Interpretations

Experimental
results

Summary



O. Bastani, C. Kim, and H. Bastani, “Interpreting blackbox models via model extraction,” *arXiv preprint arXiv:1705.08504*, 2017.