November 24th 2023, Ethical AI Workshop

CALIME

Causality-Aware Local Interpretable Model-Agnostic Explanations

Martina Cinquini
martina.cinquini@phd.unipi.it

Riccardo Guidotti
riccardo.guidotti@phd.unipi.it

Outline

1

Introduction

Lime

Causality

2

Methodology

Calime

3

Experiments

Datasets

Measures

4

Conclusions

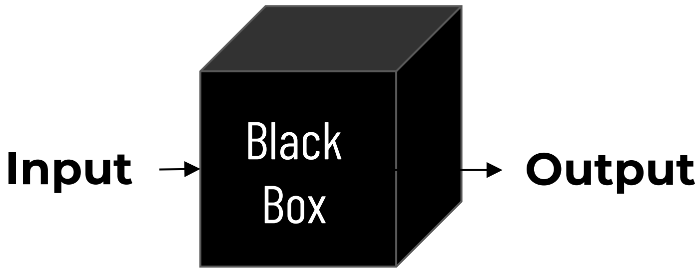
Future works

Problem

XAI approaches **do not** take into account causal relations among input features

What is eXplainable AI (XAI) ?

XAI provides explanations for the decisions of Machine Learning models.



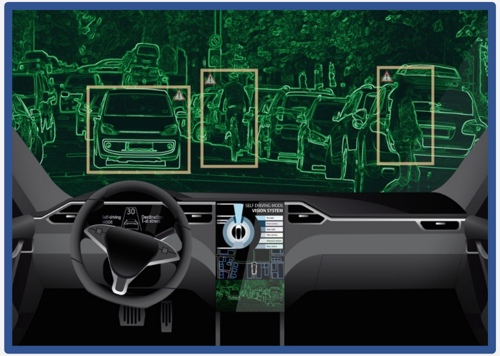
Black box models have an hidden internal structure that humans do not understand e.g. DNNs, SVMs

Source: Google Trends for "Explainable AI"

Why does XAI matter in Machine Learning?

Benefits

1. AI systems are increasingly used in sensitive areas



Self-driving cars

2. ML models can perpetuate existing bias

Racial Bias

3. Automated decision making requires reliability and trust

Financial Services

Taxonomy

Explainable by Design

Build **interpretable** ML models

Black box Explanation

Derive explanations for **complex** ML models

- Local
- Global

- Model Specific
- Model Agnostic

[1] A Survey of Methods for Explaining Black Box Models, Guidotti et al., 2018

Taxonomy

Explainable by Design

Build **interpretable** ML models

Black box Explanation

Derive explanations for **complex** ML models

Local

Global

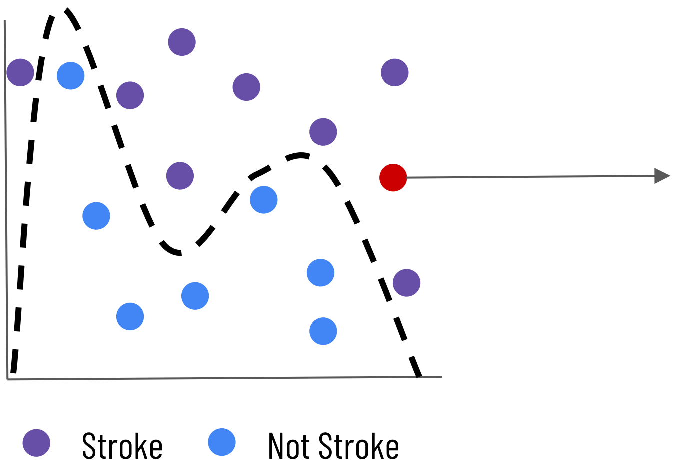
Model Specific

Model Agnostic

[1] A Survey of Methods for Explaining Black Box Models, Guidotti et al., 2018

LIME

Local **I**nterpretable **M**odel-**A**gnostic **E**xplanations[2]



Mark

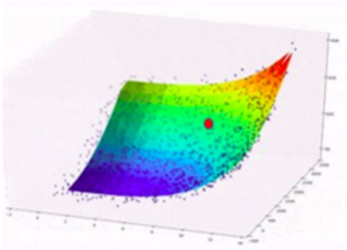
GOAL

Understand why the ML model made a certain prediction

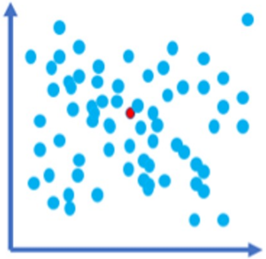
[2] "Why should I trust you?": Explaining the Predictions of Any Classifier, Ribeiro et al., 2016
Slide example from: https://www.youtube.com/watch?v=d6j6bofhj2M

LIME

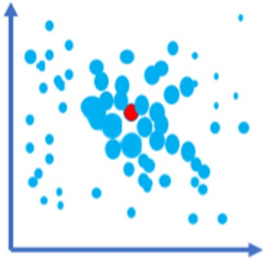
Train a black box model



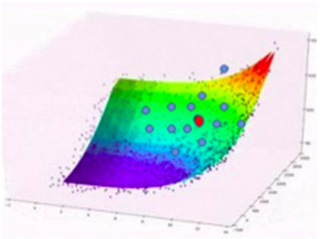
Generate random points



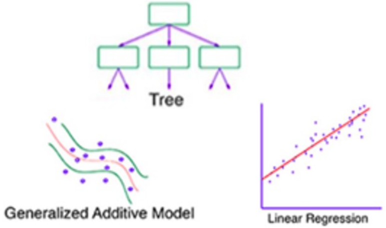
Weight based on distance



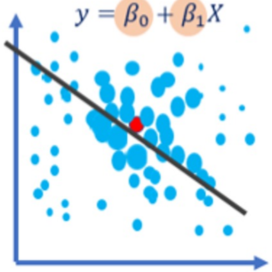
Predict the new points



Choose an interpretable model

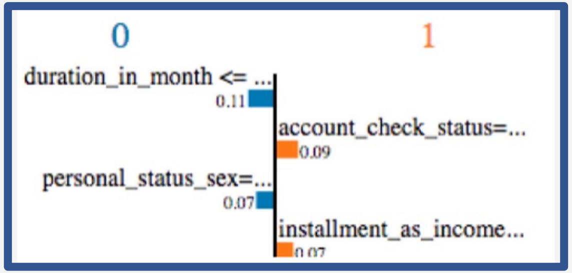


Train the model and use for explanations



LIME

Explanations

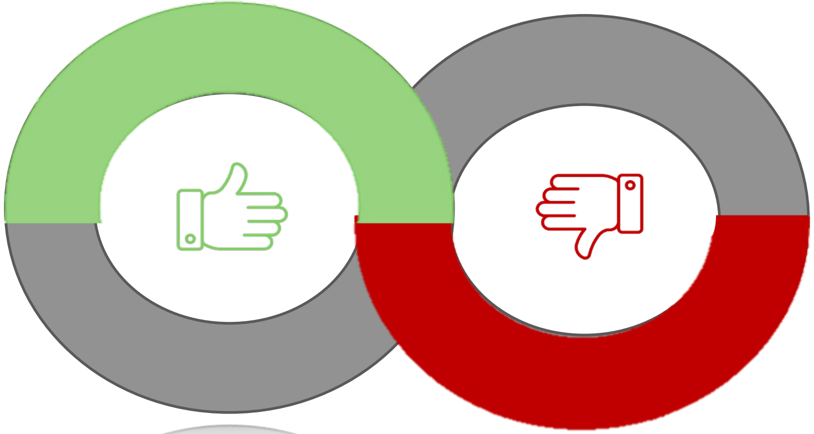


LIME

Pros & Cons

It is Model Agnostic

It works on text, images and tabular data



Instability of Explanations

Low Fidelity

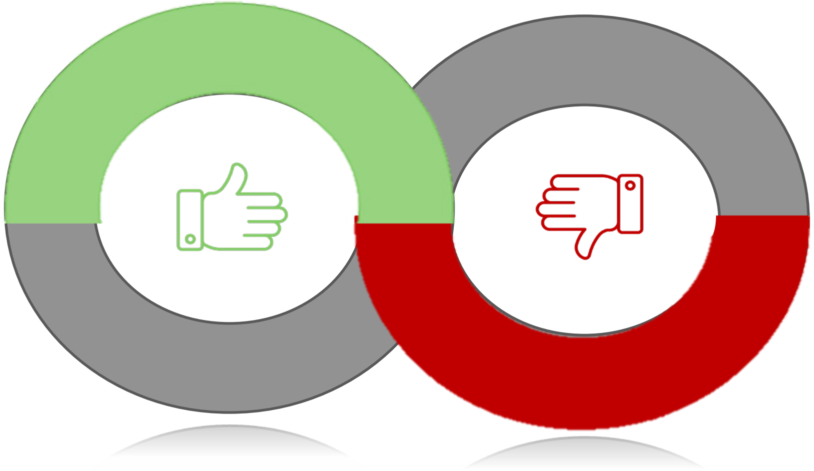
It does not consider the causal relationships among input features

LIME

Pros & Cons

It is Model Agnostic

It works on text, images and tabular data



Instability of Explanations

Low Fidelity

It does not consider the causal relationships among input features

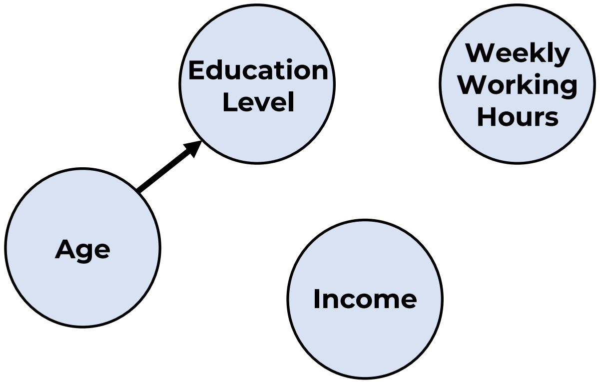
Why do we need causality?

Goal: Can the customer get the loan?

Dataset

Age	Income	Education Level	Weekly working hours
24	800	High School	20
28	1300	Bachelor Degree	35
...	...	...	...

Causal Graph



Why do we need causality?

Goal: Can the customer get the loan?

Age	Income	Education Level	Weekly working hours
24	800	High School	20
28	1300	Bachelor Degree	35
...	...	...	...

Why do we need causality?

Goal: Can the customer get the loan?

Age	Income	Education Level	Weekly working hours
24	800	High School	20
28	1300	Bachelor Degree	35
...	...	...	...

Black Box Prediction: No

Why do we need causality?

Goal: Can the customer get the loan?

Age	Income	Education Level	Weekly working hours
24	800	High School	20
28	1300	Bachelor Degree	35
...	...	...	...

Black Box Prediction: No

Lime Explanation: Low education level is mainly responsible for the denied loan

Why do we need causality?

We inspect the neighborhood generated by LIME of the instance to explain

Age	Income	Education Level	Weekly working hours
24	800	High School	20
28	1300	Bachelor Degree	35
...	...	...	...

Why do we need causality?

We inspect the neighborhood generated by LIME of the instance to explain

Age	Income	Education Level	Weekly working hours
24	800	High School	20
28	1300	Bachelor Degree	35
...	...	...	...

Generated Neighborhood

24	800	PHD	20

Why do we need causality?

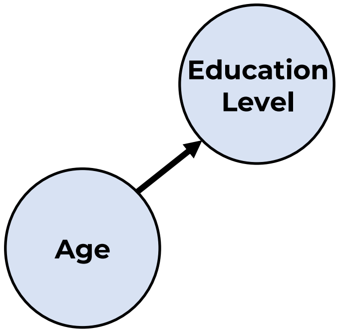
We inspect the neighborhood generated by LIME of the instance to explain

Age	Income	Education Level	Weekly working hours
24	800	**High School**	20
28	1300	Bachelor Degree	35
...	...	...	...

Generated Neighborhood

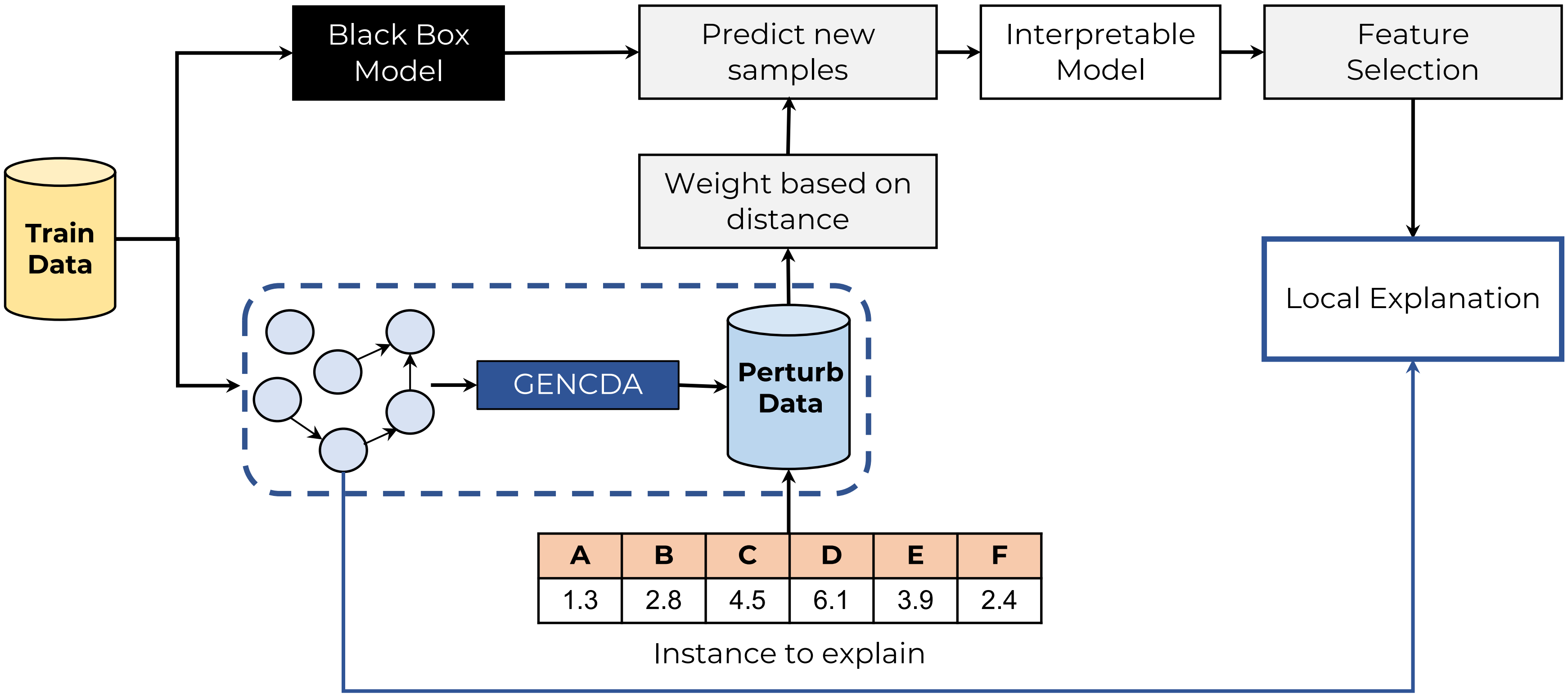
24	800	**PHD**	20

Problem: The generated instance is not plausible.
Generally, a guy who is 24 is too young to have a PhD.



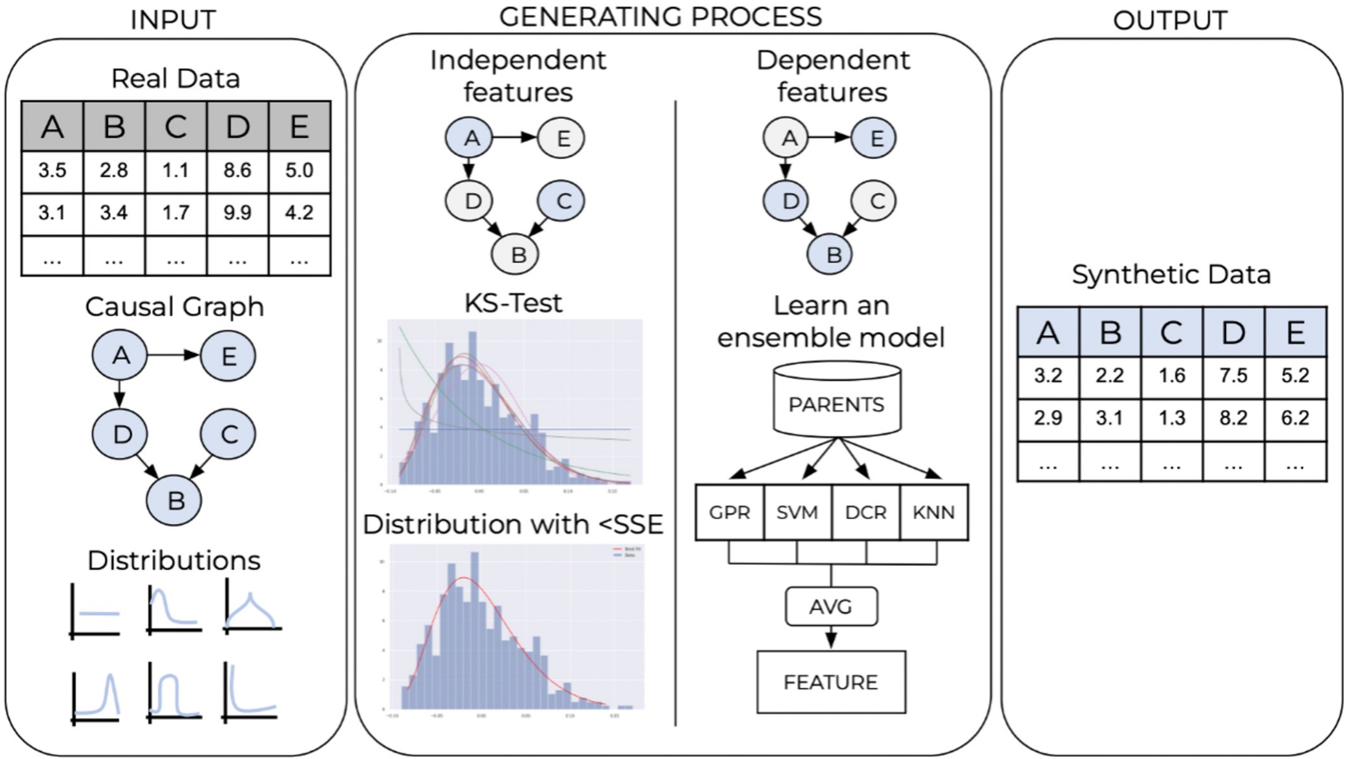
CALIME
Causality-Aware LIME

CALIME workflow



GENCDA

GEnerative **N**onlinear **C**ausal **D**iscovery with **A**priori[3]



[4] Boosting Synthetic Data Generation with Effective Nonlinear Causal Discovery, Cinquini et al., 2021

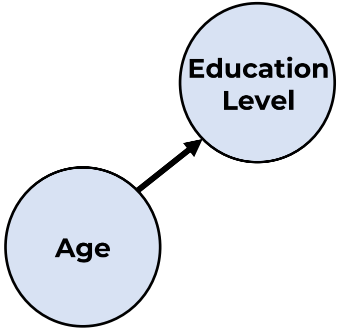
Example

We inspect the neighborhood generated by CALIME of the instance to explain

Age	Income	Education Level	Weekly working hours
24	800	**High School**	20
28	1300	Bachelor Degree	35
...	...	...	...

Generated Neighborhood

34	1500	**PHD**	30



Example

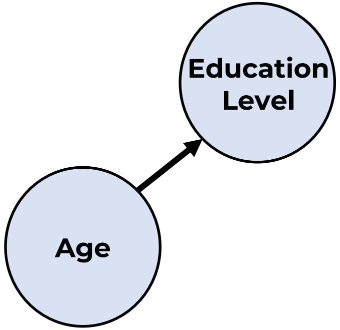
We inspect the neighborhood generated by CALIME of the instance to explain

Age	Income	Education Level	Weekly working hours
24	800	**High School**	20
28	1300	Bachelor Degree	35
...	...	...	...

Generated Neighborhood

34	1500	**PHD**	30

- Education level cannot be changed if age is not changed
- When age is changed also education level must be changed according to the regression model



Experiments

Datasets & DAGs

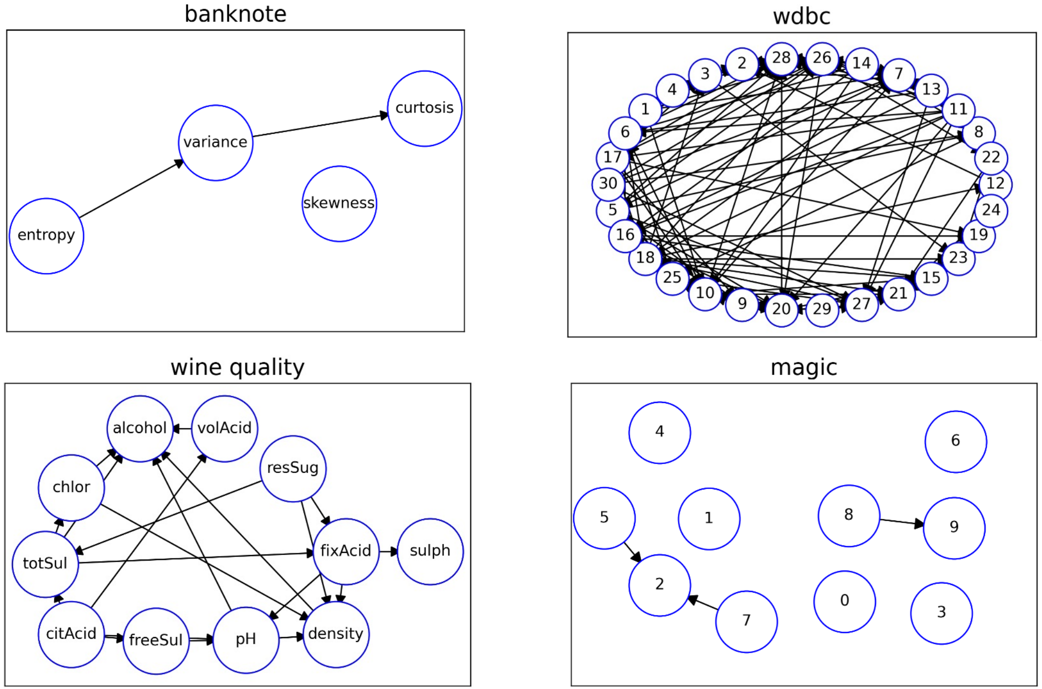
Statistics and classifiers accuracy

	n	m	RF	NN
banknote	1372	4	0.99	1.0
magic	19020	11	0.92	0.85
wdbc	569	30	0.95	0.92
wine-red	1159	11	0.82	0.70

n: # samples

m: # features

DAGs discovered by CALIME



[4] Source: UCI Repository

Evaluation Measures

Fidelity

How well does the explanation approximate the prediction of the black box model?

Plausibility

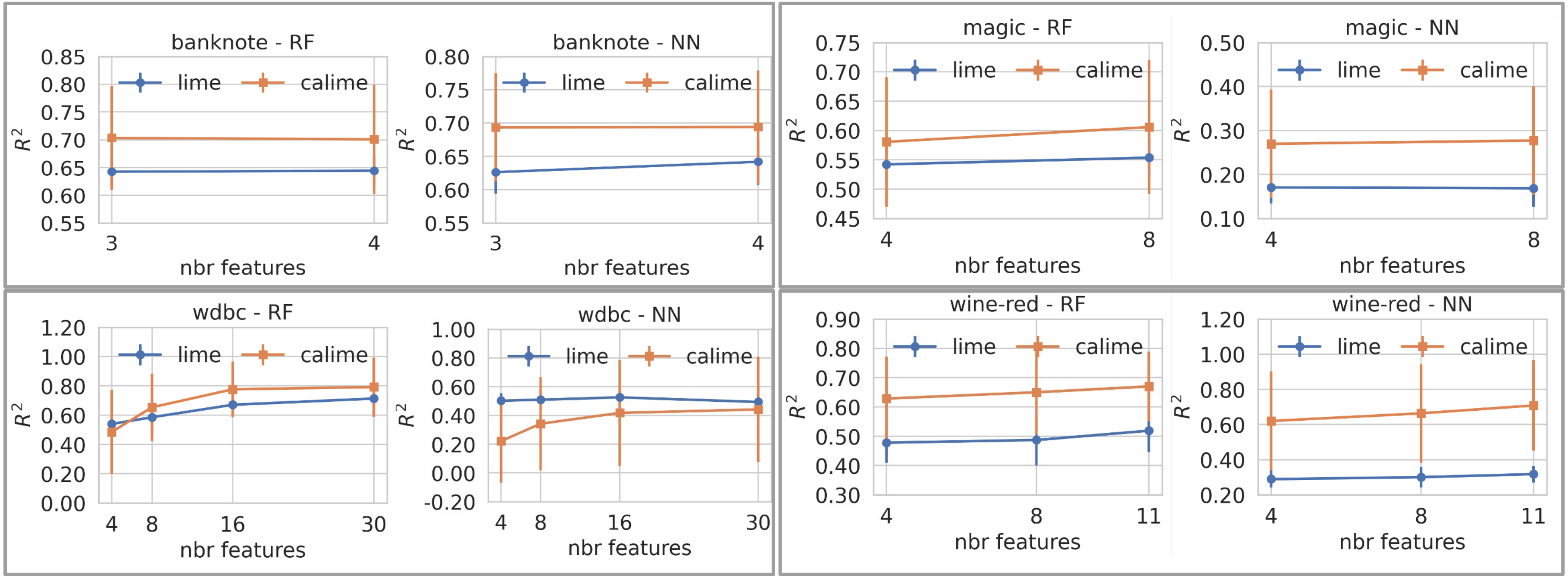
How convincing the explanations are to humans?

Stability

How similar are the explanations for similar instances?

Fidelity

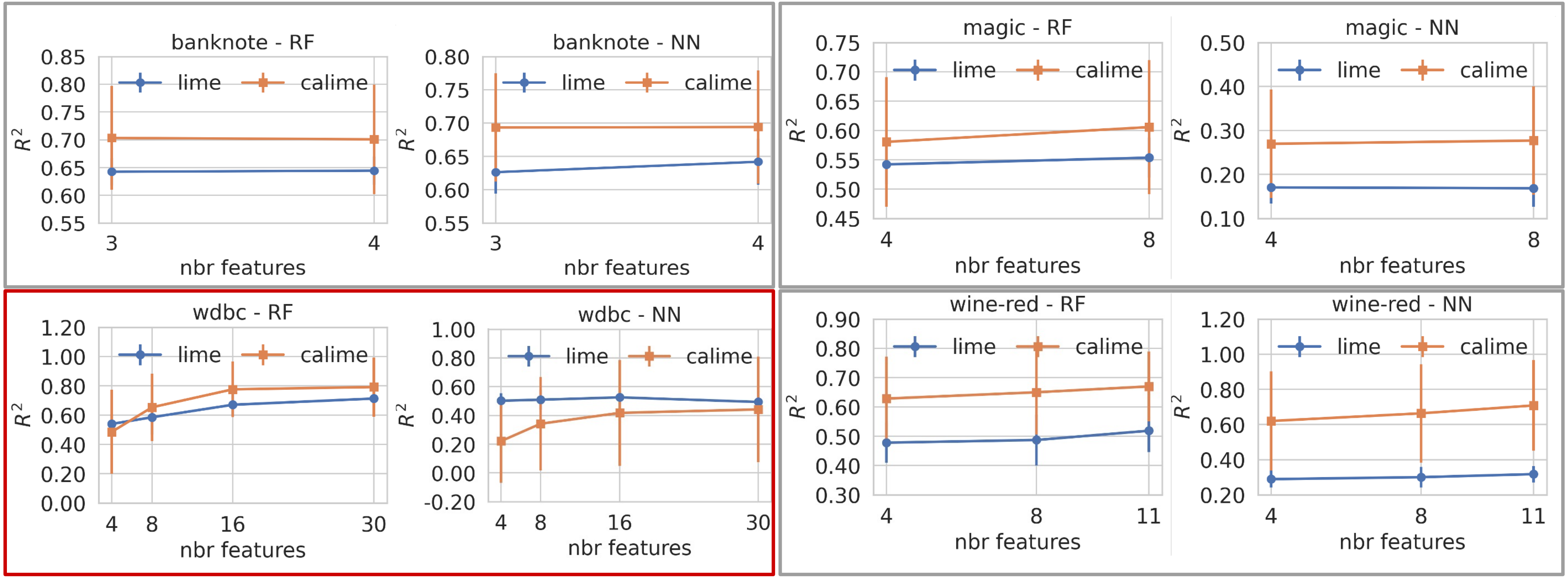
How well does the explanation approximate the prediction of the black box model?



A higher score indicates better fidelity values

Fidelity

How well does the explanation approximate the prediction of the black box model?

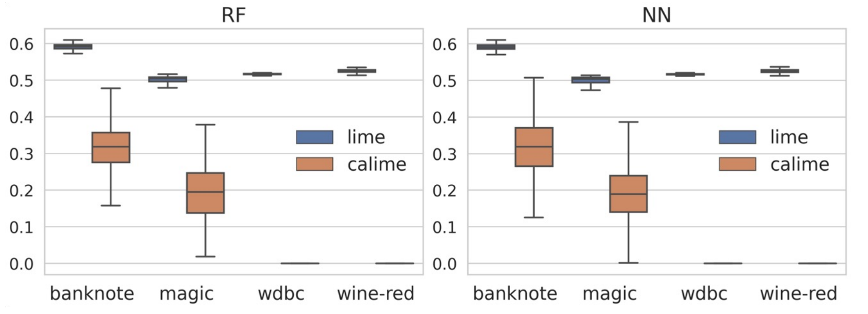


A higher score indicates better fidelity values

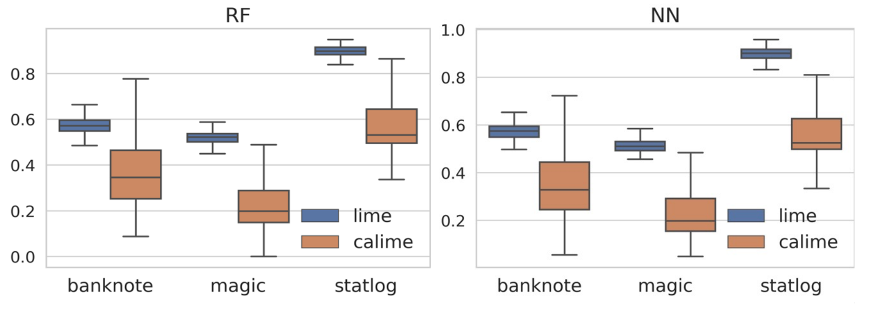
Plausibility

How convincing the explanations are to humans?

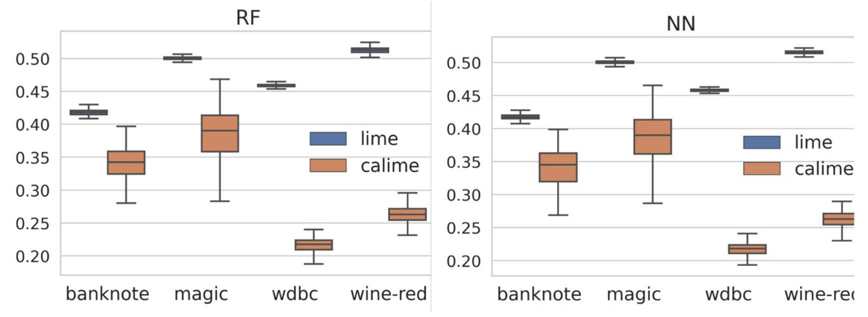
Average Minimum Distance



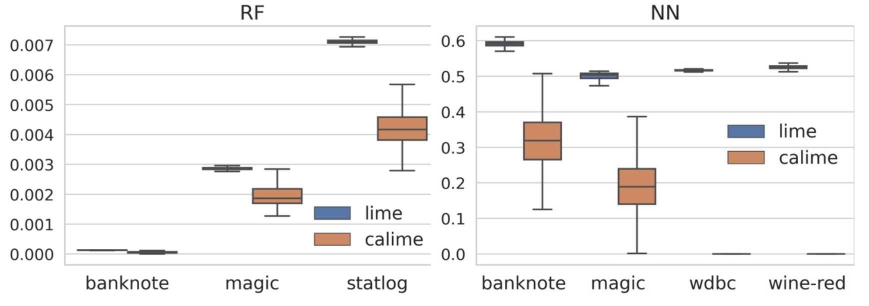
Average Outlier Score



Average Statistical Metric

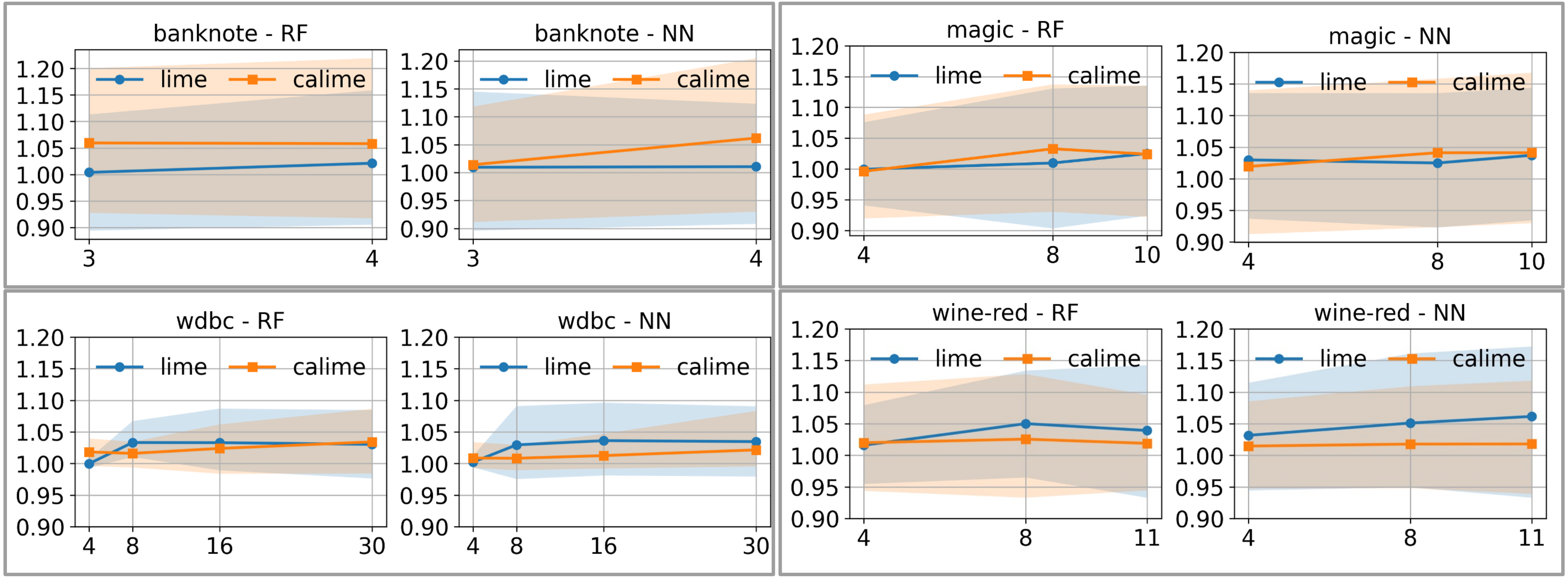


Average Detection Metrics



Stability

How similar are the explanations for similar instances?



The lower the LLE, the higher the stability.

Key takeaways

CALIME is the first black-box explanation methods returning features importance as explanations that directly discover and incorporate causal relationships in the explanation extraction process.

Key takeaways

CALIME is the first black-box explanation methods returning features importance as explanations that directly discover and incorporate causal relationships in the explanation extraction process.

Experiments results show that CALIME overcomes the weaknesses of LIME concerning both the fidelity in mimicking the black-box and the stability of the explanations.

Key takeaways

CALIME is the first black-box explanation methods returning features importance as explanations that directly discover and incorporate causal relationships in the explanation extraction process.

Experiments results show that CALIME overcomes the weaknesses of LIME concerning both the fidelity in mimicking the black-box and the stability of the explanations.

CALIME could strengthen user trust in the AI system. It will be especially useful for high-impact domains such as financial services or healthcare (e.g., therapy planning or patient monitoring).

Key takeaways

Ethical AI:

- Transparency through causal explanations helps mitigate concerns related to algorithmic bias and unfairness, contributing to a more trustworthy AI ecosystem.

Key takeaways

Ethical AI:

- Transparency through causal explanations helps mitigate concerns related to algorithmic bias and unfairness, contributing to a more trustworthy AI ecosystem.

Future Directions:

- Develop causality aware explanation methods suitable for images and time series working in a similar manner of CALIME;

- Employ the knowledge about causal relationships in the explanation extraction process of other model-agnostic explainers like LORE, SHAP or ANCHORS.

Thank you for your attention!

GitHub /marti5ini

in /martinacinquini

http://pages.di.unipi.it/cinquini/