Explainable Artificial Intelligence (XAI)

Azqa Nadeem

PhD candidate @ Cyber Analytics Lab

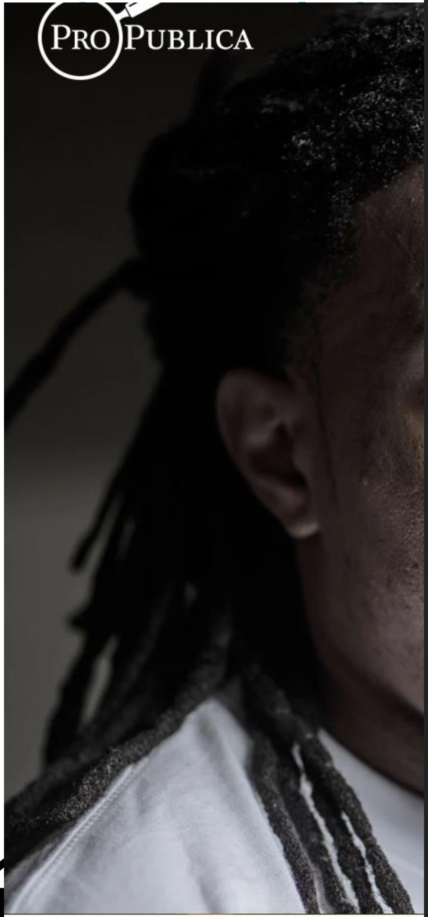
Department of Intelligent Systems

Delft University of Technology

TU Delft

16 December 2022

Cyber Analytics Lab


ProPublica


Two Shoplifting Arrests
JAMES RIVELLI
Prior Offenses
1 domestic violence aggravated assault, 1 grand theft, 1 petty theft, 1 drug trafficking
ROBERT CANNON
Prior Offense
1 petty theft

https://www.pro

White-box vs Black-box classifiers

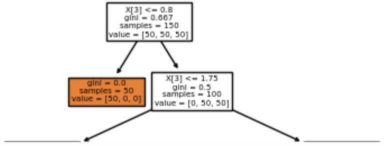
- *A white-box classifier is transparent in terms of the function it represents, and can thus be understood by human experts.*

- *A black-box classifier often aims for optimal performance at the cost of interpretability, i.e., they represent a function that is difficult for human experts to understand.*

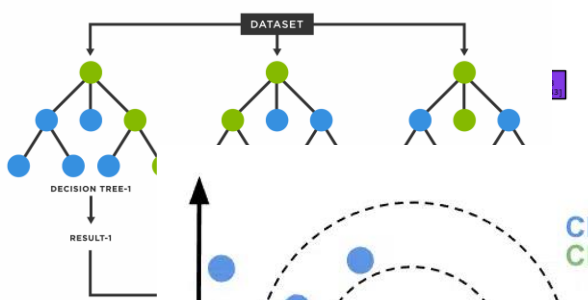
- A (relatively) simple test: given inputs and outputs, can a human interpret the relationship between them?

Are these white-box or black-box?

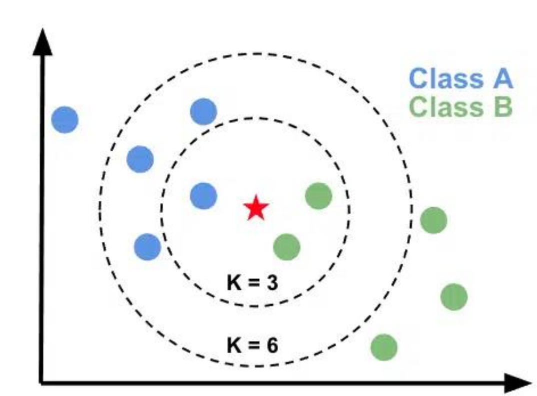
- Decision tree



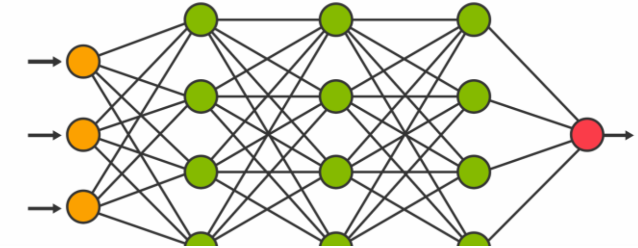
- Randc



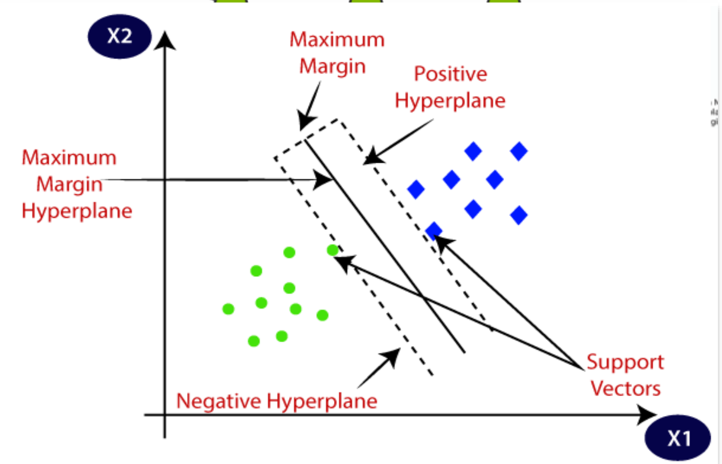
- K-nea



- Neural networks



- Suppor



https://scil
https://ww
https://ww
https://ww
https://ww
hine-algorithm

TU Delft

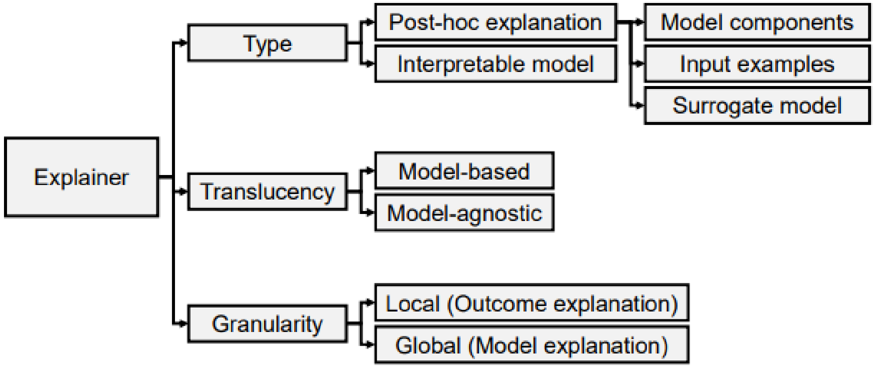
Explainable Artificial Intelligence (XAI)

- *"XAI provides a set of tools and techniques that aim to make machine learning models <u>human understandable</u> by explaining either the <u>model predictions and/or the input data</u>."*

- Interpretable vs. explainable
  - White-box model vs. explaining an ML model

- Model-based vs. model-free
  - Whether the explanation method works with a specific model

- Local vs. global explanations
  - Explaining a single vs. all data instances



TU Delft

Nadeem, Azqa, et al. "Sok: Explainable machine learning for computer security applications." arXiv preprint arXiv:2208.10605 (2022).

What is an explanation? (…in AI)

Explanation contains a causal chain and explanation selection.

An explainee cares only about a subset of causes w.r.t. their context. From those, the explainer may select a few causes, and the explainer and explainee may interact about them.

Although, explanations are often restricted to causal attribution in AI…

TU Delft

Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." Artificial intelligence 267 (2019): 1-38.

Properties of good explanations

- Explanations are selected (from many causes)
  - Select a few (biased) causes from an exhaustive list
- Explanations are social
  - Transfer of knowledge; tailored to explainers' beliefs about explainee's beliefs
- Explanations are contrastive
  - **Why e happened?** vs. **Why e happened instead of x?**
- Referring to causes is more effective than probabilities
  - The most likely explanation is not necessarily the best one for the explainee

TUDelft

Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." Artificial intelligence 267 (2019): 1-38.

Properties of good explanations

- Explanations are selected (from many causes)
  - Select a few (biased) causes from an exhaustive list
- Explanations are social
  -
- Exp
  -

Good explanations are ones that an explainee will actually use. User studies are an important part of evaluating the usefulness of explanations!

- Referring to causes is more effective than probabilities
  - The most likely explanation is not necessarily the best one for the explainee

Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." Artificial intelligence 267 (2019): 1-38.

How to explain ML models?

Dataset used in this lecture…

- Cervical cancer (risk factors) prediction dataset, UCI ML repo
- 858 rows, 35 features, 1 target label (healthy/cancer)
- Base model: Decision tree (White-box), Random Forest (Black-box)

Attribute Information:

(int) Age
(int) Number of sexual partners
(int) First sexual intercourse (age)
(int) Num of pregnancies
(bool) Smokes
(bool) Smokes (years)
(bool) Smokes (packs/year)
(bool) Hormonal Contraceptives
(int) Hormonal Contraceptives (years)
(bool) IUD
(int) IUD (years)
(bool) STDs
(int) STDs (number)
(int) STDs: Number of diagnosis
(int) STDs: Time since first diagnosis
(int) STDs: Time since last diagnosis
(bool) Dx:Cancer
(bool) Dx:CIN
(bool) Dx:HPV
(bool) Dx
(bool) Hinselmann: target variable
(bool) Schiller: target variable
(bool) Cytology: target variable
(bool) Biopsy: target variable

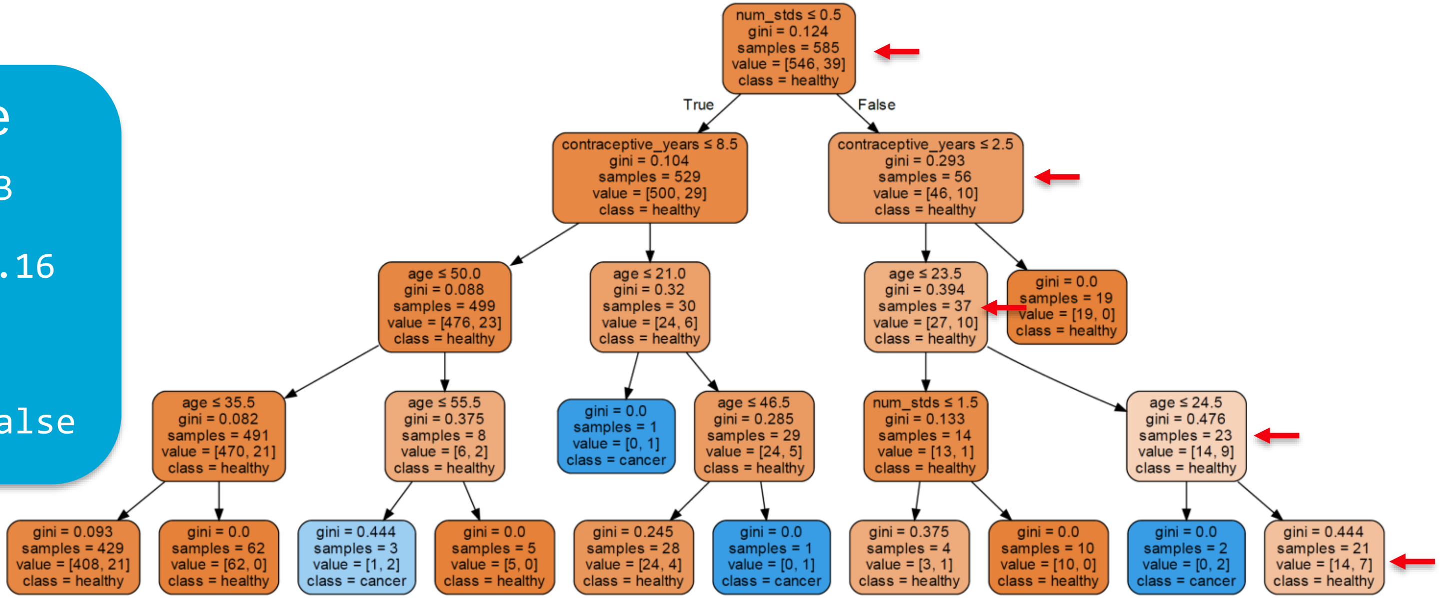
https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29

Decision tree [Local/Global] [Interpretable] [Model-based]

- Explains the dataset globally, and explains single instances by tracing a tree path

Test instance

age	33
contra_years	0.16
num_stds	1
smokes	False



TU Delft

Decision tree - Analysis

- Creates local and global explanations
- Can validate the model directly

- May not be the most accurate model for the task

TU Delft

Permutation Importance [Global] [Post-hoc] [Model-agnostic]

- Explains the impact of permuting a feature on the classifier loss, breaking the relationship between the feature and true outcome
  - High loss discrepancy → important feature
  - Low loss discrepancy → unimportant feature
- Repeat multiple times and average out the loss discrepancy

Age	contra_years	num_std	smokes	Label
37	0.25	2	0	0
19	0.5	0	0	0
18	0	0	0	1

Loss = 0.11

Age

$Loss_{age} = 0.29$

Δ loss = 2.6

Num_std

$Loss_{std} = 0.19$

Δ loss = 1.7

TU Delft

Permutation Importance

- Explains the impact of permuting a feature on the classifier loss, breaking the relationship between the feature and true outcome
  - High loss discrepancy → important feature
  - Low loss discrepancy → unimportant feature
- Repeat multiple times and average out the loss discrepancy
- Compute on test data!

Training data

age	0.044 +/- 0.005
contraceptive_years	0.044 +/- 0.005
num_stds	0.026 +/- 0.004
smokes	0.012 +/- 0.003

Test data

smokes	0.003 +/- 0.006
num_stds	-0.003 +/- 0.007
age	-0.008 +/- 0.009
contraceptive_years	-0.013 +/- 0.007

Evidence of overfitting!

Permutation Importance - Analysis

- Detect features that hurt the generalizability of the model
- Can be used to explain any black-box model

- Directly linked to the loss of a model
  - Not necessarily marginal contribution of a feature for a given prediction
- Tricky interpretation with correlated features
  - Loss discrepancy include main feature effect & interaction effects
  - Generates impossible data instances while permutation
  - Underestimates importance of correlated features

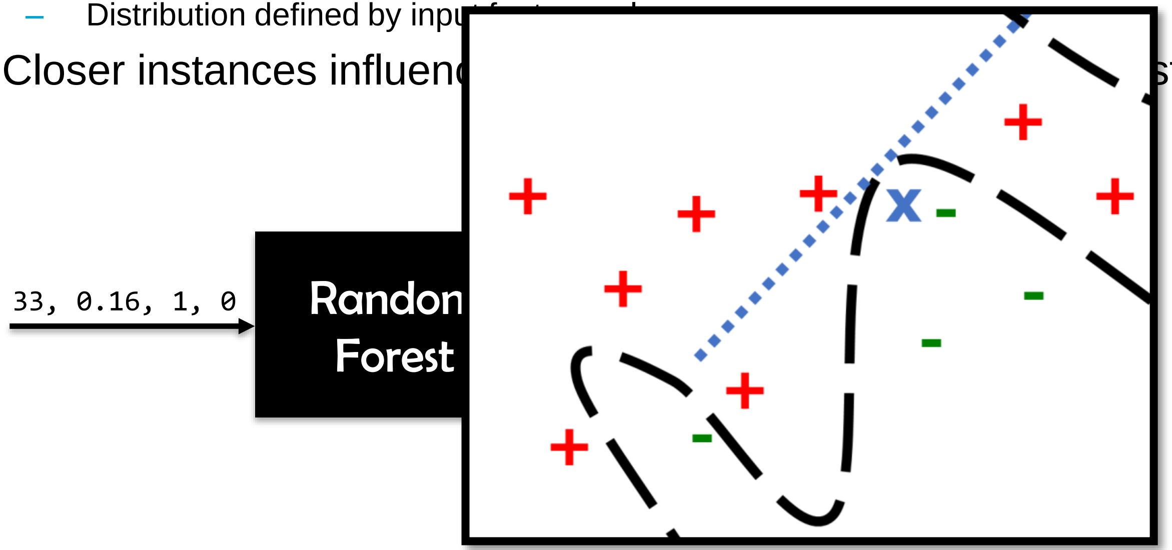
LIME [Local] [Post-hoc] [Model-agnostic]

- Explains a prediction by learning a local surrogate model for the data instance
  - Approximates the predictions of the black-box model in a local neighborhood
- Input instance perturbed for each feature by sampling from a normal distribution
  - Distribution defined by input feature values
- Closer instances influence the surrogate more than farther instances

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016.

LIME [Local] [Post-hoc] [Model-agnostic]

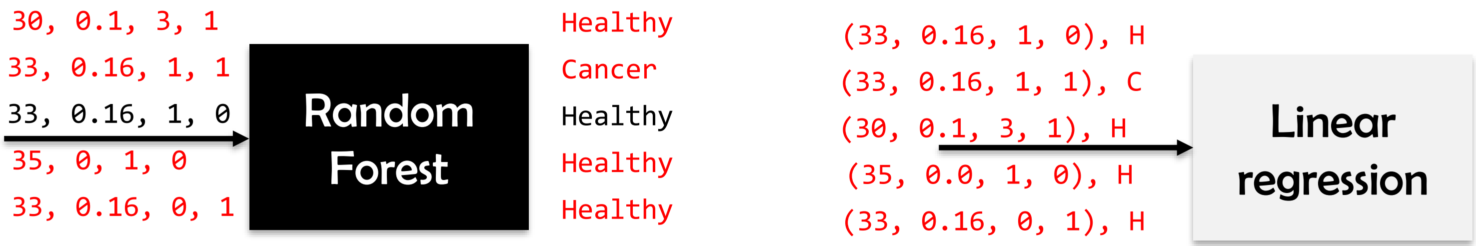
- Explains a prediction by learning a local surrogate model for the data instance
  - Approximates the predictions of the black-box model in a local neighborhood
- Input instance perturbed for each feature by sampling from a normal distribution
  - Distribution defined by inpu
- Closer instances influenc … stances



Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016.

LIME [Local] [Post-hoc] [Model-agnostic]

- Explains a prediction by learning a local surrogate model for the data instance
  - Approximates the predictions of the black-box model in a local neighborhood
- Input instance perturbed for each feature by sampling from a normal distribution
  - Distribution defined by input feature values
- Closer instances influence the surrogate more than farther instances



TU Delft

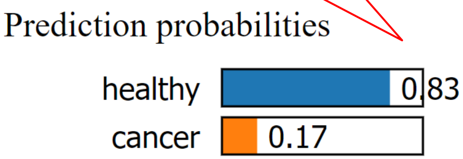
Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016.

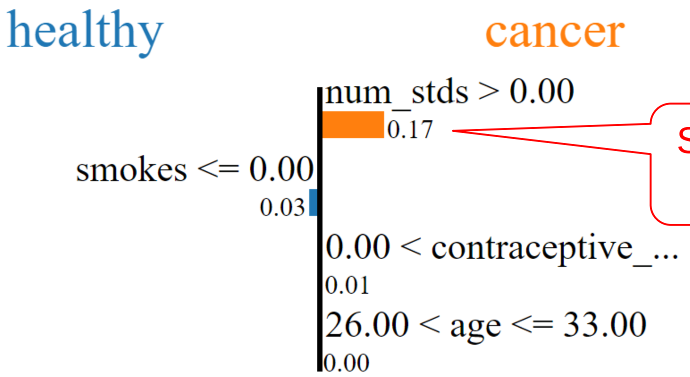
LIME

(33, 0.16, 1, 0), H
(33, 0.16, 1, 1), C
(30, 0.1, 3, 1), H
(35, 0.0, 1, 0), H
(33, 0.16, 0, 1), H

Linear regression

RF prediction probabilities





Surrogate weight * feature value

Feature	Value
num_stds	1.00
smokes	0.00
contraceptive_years	0.16
age	33.00

Features used in explanation generation

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016.

LIME - Analysis

- Free choice of local interpretable model
- Limit the features used for explanation generation

- Unclear from the explanations how far one can extrapolate from the predictions
  - How big should the local neighborhood be?
  - Changes the explanations dramatically
- Explanations may change for different sampling runs

Counterfactual [Local] [Post-hoc] [Model-agnostic]

- Explains the minimal feature change(s) that alter the prediction for an instance
  - Similar to the original instance
  - Change minimal features possible
  - Changes to feature values must be realistic

- Has a causal form: “If X had not happened, Y would also not have happened”
  - E.g., “The shop is closed either because it is raining or the owner is sick.”
  - Counterfactual: “The shop is open because it is raining and the owner is healthy.”

	Age	contra_years	num_std	smokes	Label
	20	0.25	0	0	0
CF#1	25	0.25	1	0	1
CF#2	20	2	4	0	1

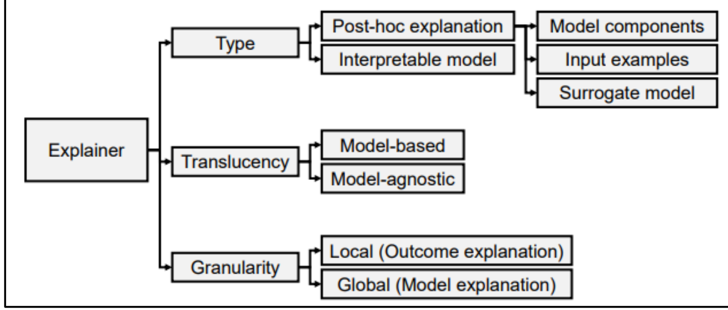
TU Delft

Counterfactual - Analysis

- Natural interpretation of counterfactual explanations
  - Report only what has changed
- Creates new (artificial) data instances as explanations
- Expensive to create counterfactuals that fulfill all constraints
- Rashomon effect: Multiple contradictory counterfactual explanations can exist
  - Which one to report?

TUDelft

Recap



Decision tree

- Interpretable
- Model-based
- Global/local

Permutation importance

- Post-hoc (feature summary)
- Model-agnostic
- Global

LIME

- Post-hoc (surrogate)
- Model-agnostic
- Local

Counterfactuals

- Post-hoc (example-based)
- Model-agnostic
- Local

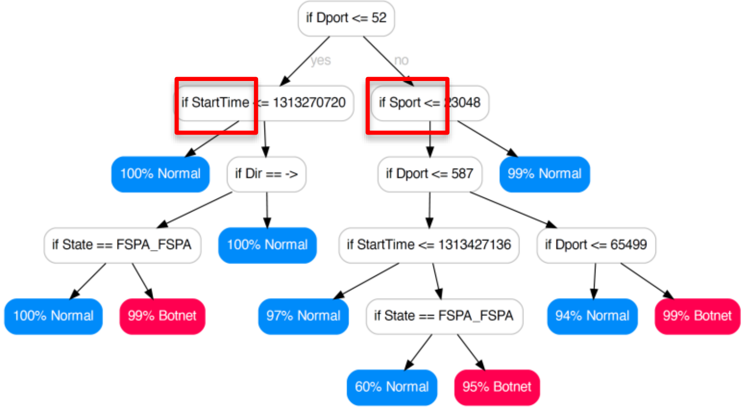
TU Delft

Examples: XAI in cybersecurity

Debugging a malicious network traffic detector

- Gradient Boosting Machine learnt on Netflow data
  - Tabular features: start time, duration, protocol, source port, …
  - Binary classification task: Normal | Botnet
  - Balanced accuracy: 86.4%

- Q1. Does the model use the correct features?
  - Interpretable decision tree shows problematic features
  - *Solution: retrain without spurious features for better generalizability*
    - *Balance accuracy drops to 74.4%*



TU Delft

Nadeem, Azqa, et al. "Sok: Explainable machine learning for computer security applications." arXiv preprint arXiv:2208.10605 (2022).

Debugging a malicious network traffic detector

- Gradient Boosting Machine learnt on Netflow data
  - Tabular features: start time, duration, protocol, source port, …
  - Binary classification task: Normal | Botnet
  - Balanced accuracy: 86.4%

- Q2. Where does the GBM make mistakes?
  - LIME shows a false negative (missed malicious flow)
  - dport suggests Netflow is benign
  - *Solution: Fix the experimental dataset or learn from a more realistic dataset*

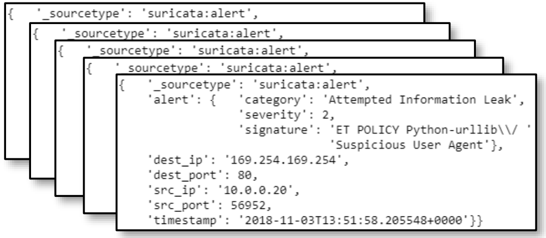
Feature	Value	LIME Rule	Weight
Dport	3389	Dport = 3389	0.18
StartTime	1313571534	1313537772.00 < Start...	0.13
Sport	4505	Sport=4505	0.09
TotPkts	10	TotPkts > 4.00	0.07
State	16	State=16	0.04
Proto	0	Proto=0	0.03
SrcBytes	437	SrcBytes > 186.0	0.03
TotBytes	1076	TotBytes > 494.25	0.02
Dir	2	Dir = 2	0.01
Dur	60.95	Duration > 9.01	0.01

TU Delft

Nadeem, Azqa, et al. "Sok: Explainable machine learning for computer security applications." arXiv preprint arXiv:2208.10605 (2022).

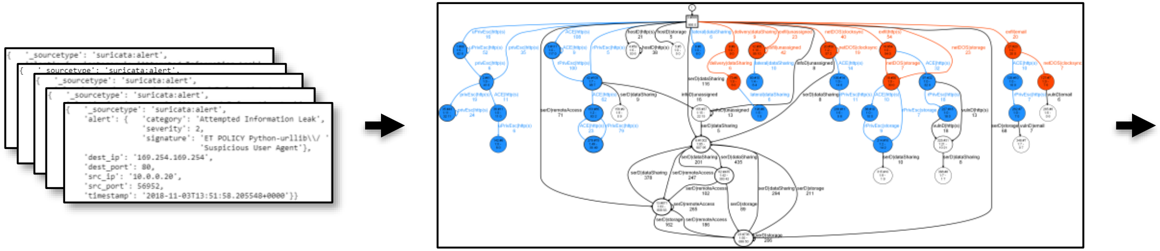
Extracting attacker strategies from intrusion alerts

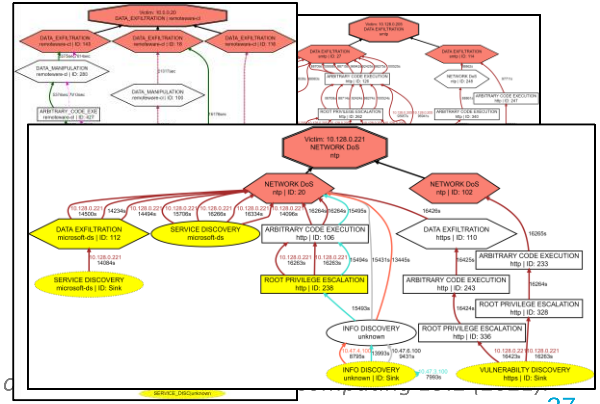
- Security analysts are overloaded with intrusion alert investigation



1 million alerts/day!

- Q3. What can we learn about attacker strategies by analyzing alerts?





Nadeem, Azqa, et al. "Alert-driven attack graph generation using s-pdfa." IEEE Transactions o... 731-746.

Summary

- XAI aims to explain the black-box model predictions or input data
  - For usability, verification and establishing trust
- Good explanations are
  - Contrastive, selected, social, and tailored to the explainee
- A few explanation methods for tabular data
  - Decision tree for interpretable ML
  - Permutation Importance for feature summary
  - LIME for local linear surrogate model
  - Counterfactuals for nearest-unlike explanations
- XAI can detect spurious features, discover reasons for misclassifications, and explain input data in human understandable way

TU Delft

Further reading

- Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." Artificial intelligence 267 (2019): 1-38.
- Doshi-Velez, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning." arXiv preprint arXiv:1702.08608 (2017).
- Molnar, Christoph. Interpretable machine learning. Lulu. com, 2020.
- Nadeem, Azqa, et al. "Sok: Explainable machine learning for computer security applications." arXiv preprint arXiv:2208.10605 (2022).

Questions?