ML4Seismic Partners Meeting 2023

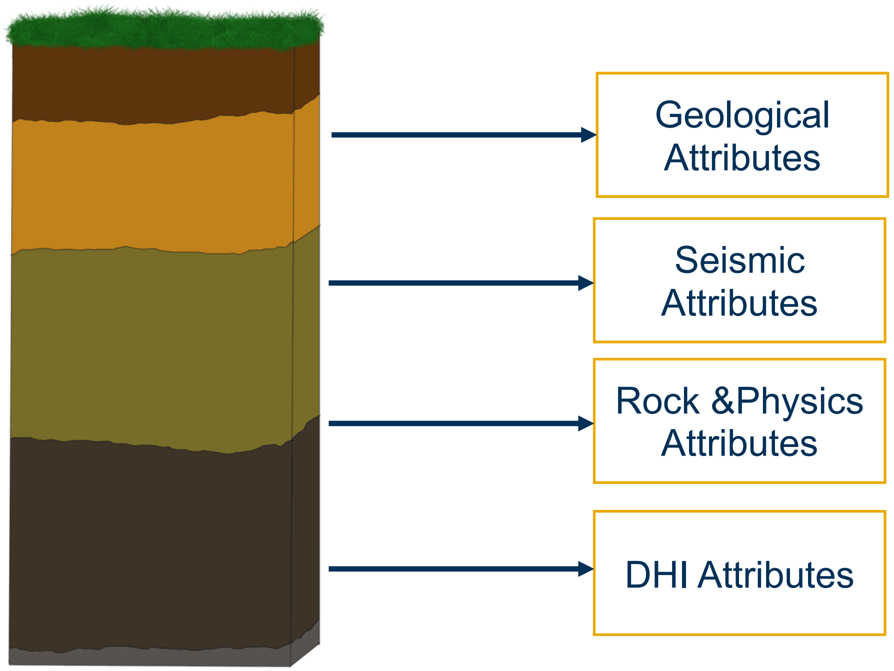
A Counterfactual Analysis of Interpretations in High Dimensional DHI Data

Prithwijit Chowdhury, Mohit Prabhushankar, Ahmad Mustafa and Ghassan AlRegib

Hydrocarbon Prospect Analysis

Dataset to access and evaluate risks associated with drilling ventures

Decision is made by incorporating different geology and geophysical attributes into a calibration system

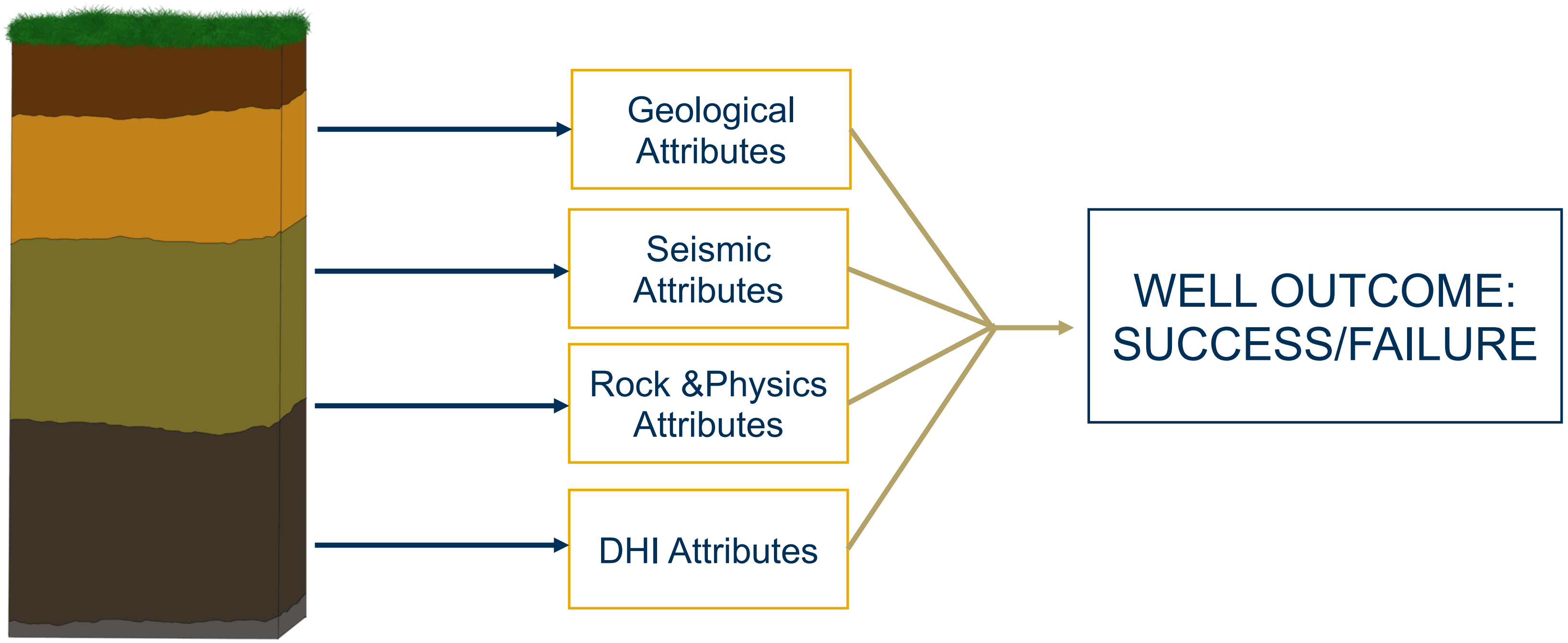


Attribute collection

Hydrocarbon Prospect Analysis

Dataset to access and evaluate risks associated with drilling ventures

Decision is made by incorporating different geology and geophysical attributes into a calibration system

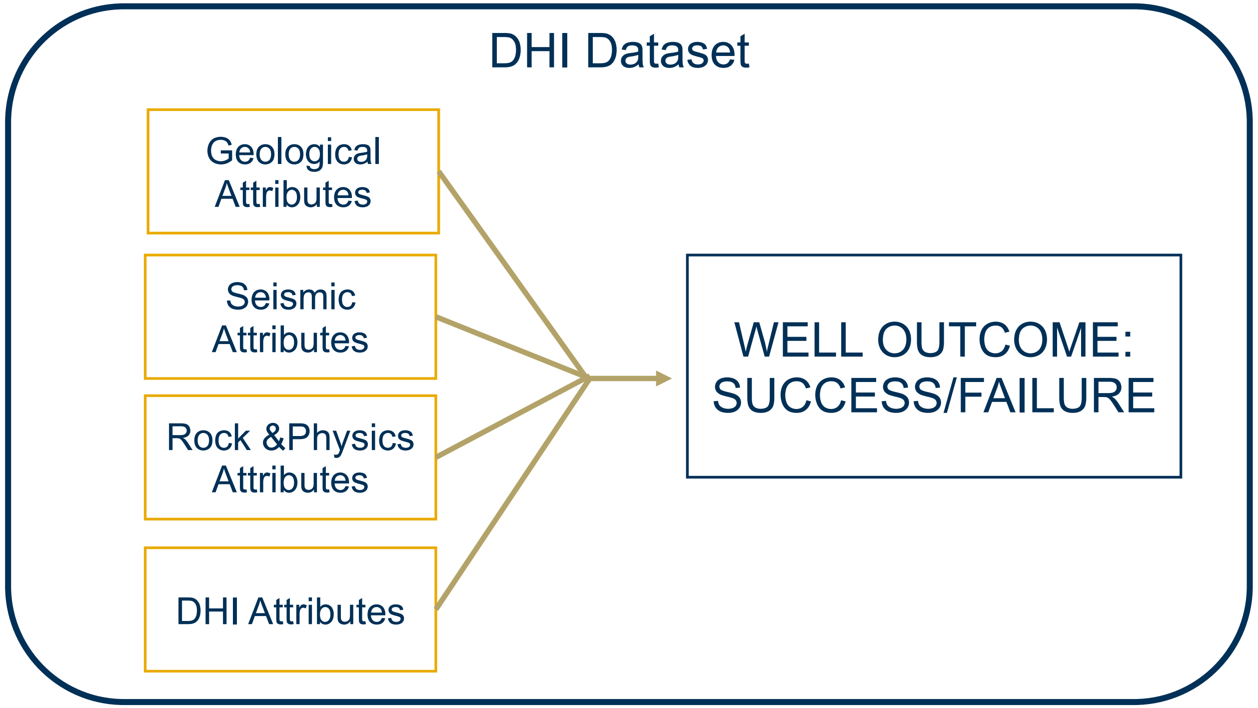


Attribute collection

Prospect risking

Direct Hydrocarbon Indicator (DHI) Dataset
Dataset to access and evaluate risks associated with drilling ventures

All the collected attributes and the final decisions made by experts are gathered into a classification dataset

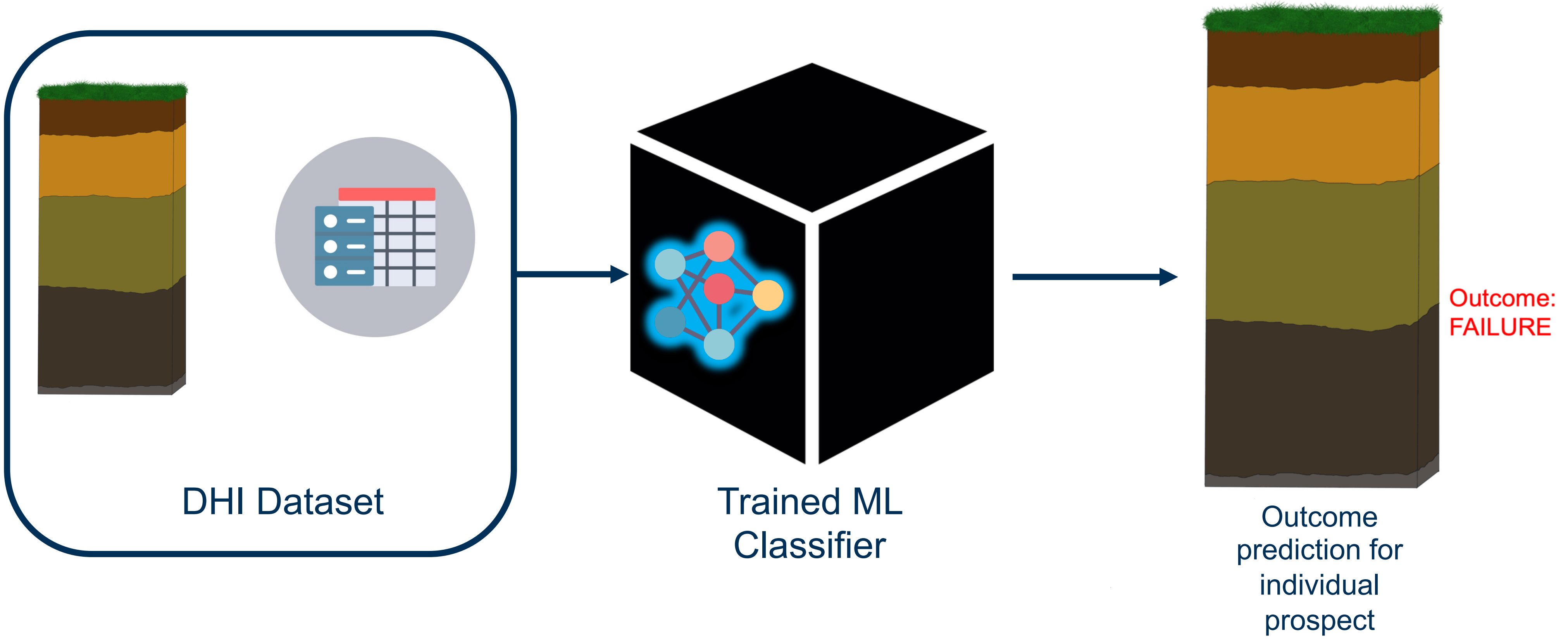


- Dataset has **33 features** (attributes)
- **350 individual prospects.**
- **Final Decision** is a binary classification for **successful (1) or failed (0) prospects**

Machine Learning (ML) based Prospect Analysis

ML Classifiers provide fast and efficient decisions

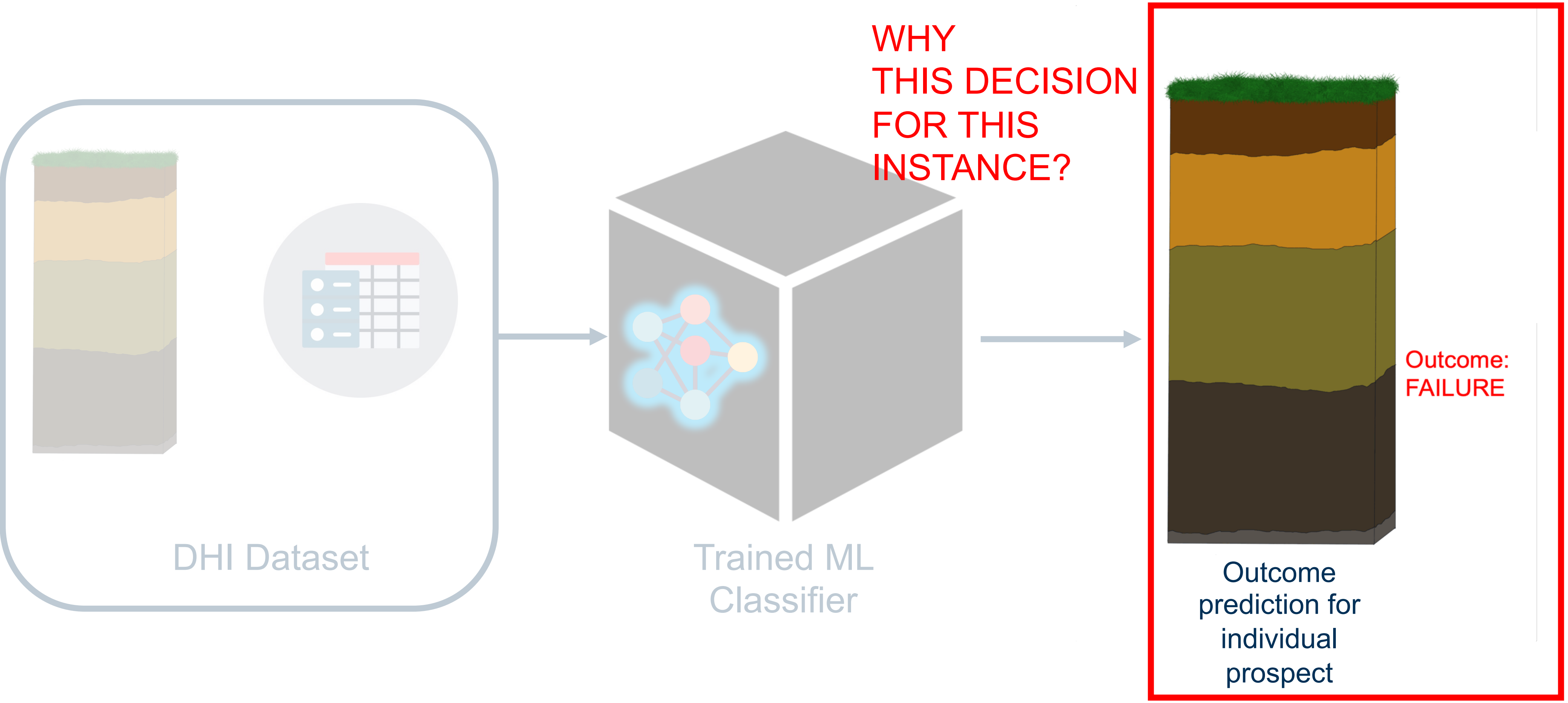
A binary classifier fitted on the DHI dataset can be used to infer decisions on prospects



Machine Learning (ML) based Prospect Analysis

ML Classifiers provide fast and efficient decisions

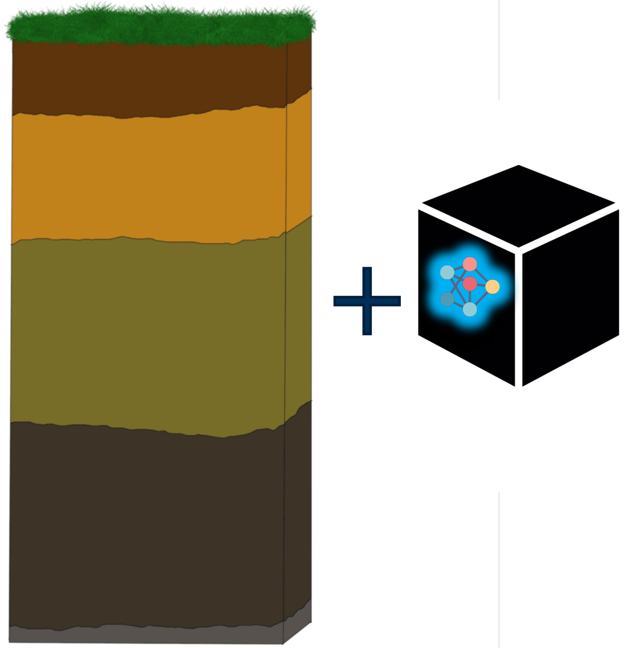
A binary classifier fitted on the DHI dataset can be used to infer decisions on prospects



Attribute based local explainable AI (XAI) methods: LIME & SHAP

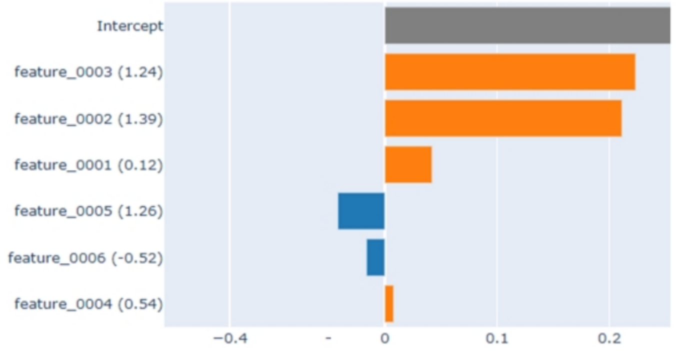
Individual decisions by the model can be further studies by observing the local explanations

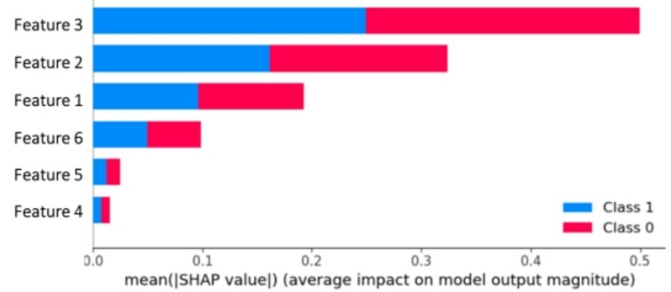
Local explanation methods explain model predictions for specific data points by ranking the input features based on importance.



DATA + MODEL PREDICTION

LOCAL XAI MODULES



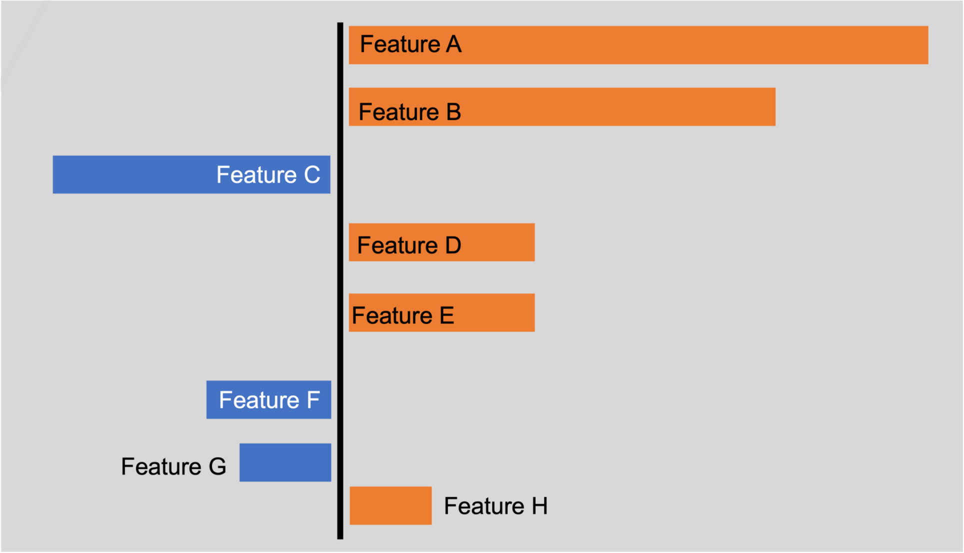


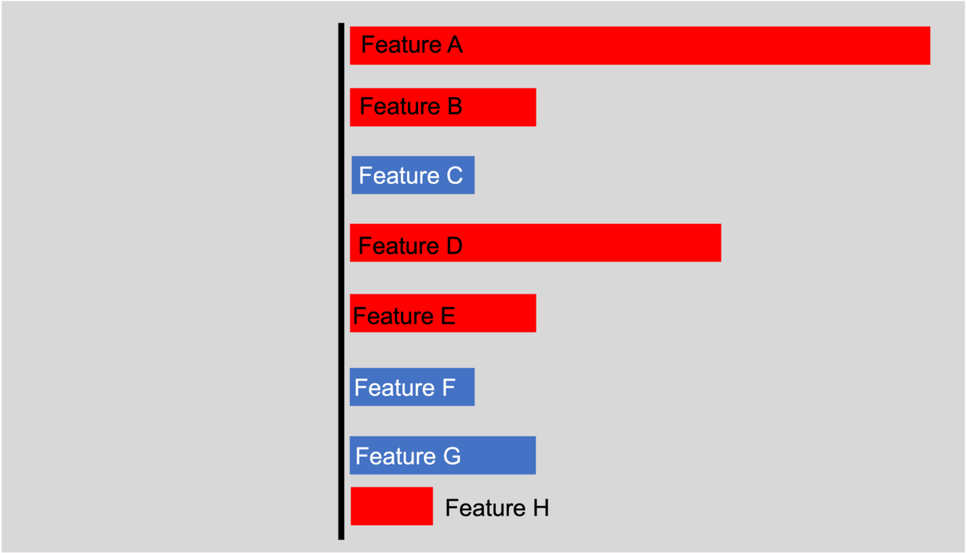
FEATURE IMPORTANCE SCORES AS EXPLANATION FOR A PREDCITION

Drawback of LIME and SHAP

These feature ranking methods suffer from a major drawback: Disagreement

Disagreement between LIME and SHAP for the same explanation





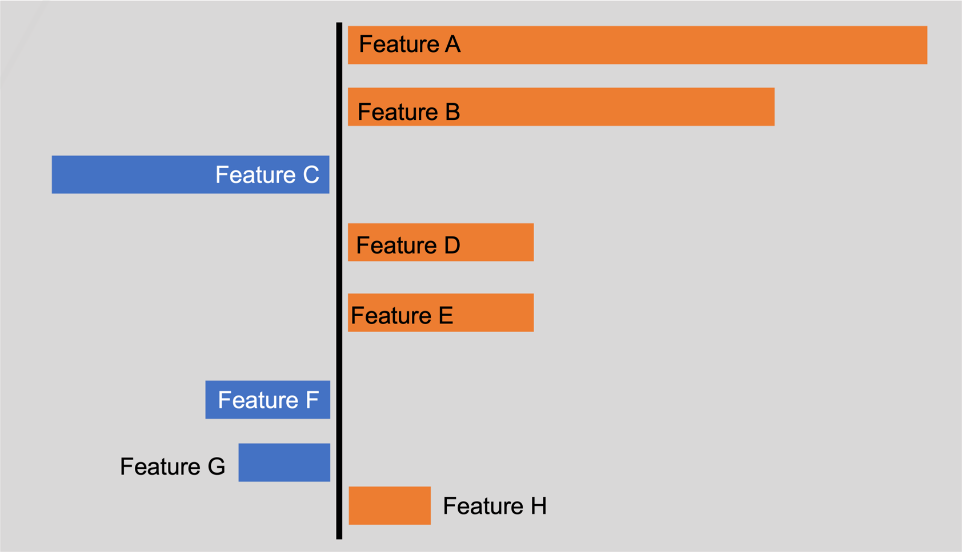
Doshi-Velez, F., and B. Kim, 2017, Towards a rigorous science of interpretable machine learning: arXiv preprint arXiv:1702.08608.

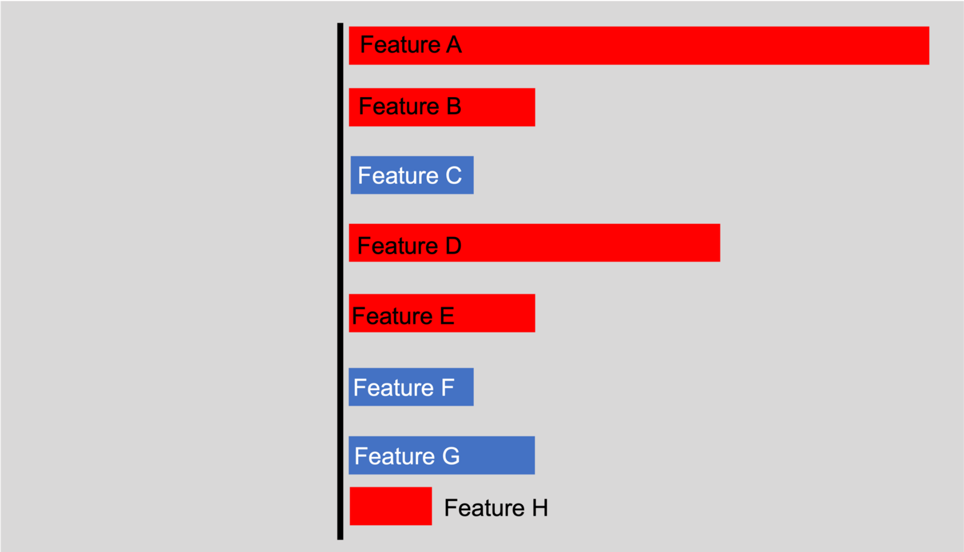
*toy examples

Drawback of LIME and SHAP

These feature ranking methods suffer from a major drawback: Disagreement

Disagreement between LIME and SHAP for the same explanation





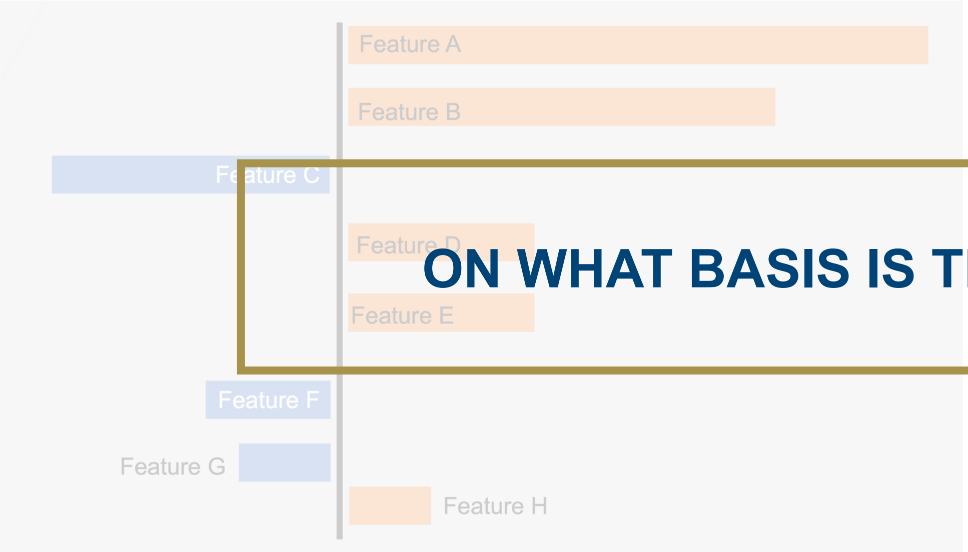
Definition of “importance” and “relevance” is different for different explainers.

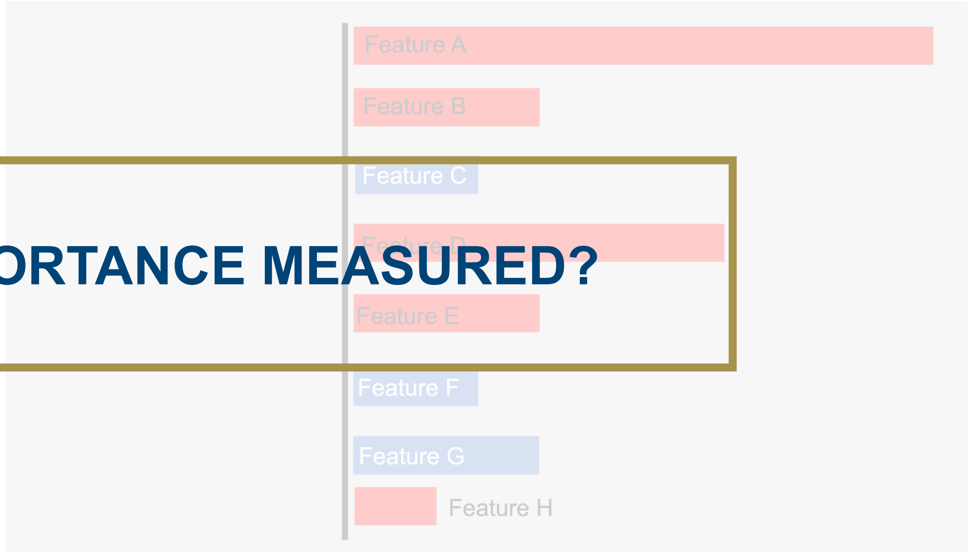
*toy examples

Drawback of LIME and SHAP

These feature ranking methods suffer from a major drawback: Disagreement

Disagreement between LIME and SHAP for the same explanation





ON WHAT BASIS IS THIS IMPORTANCE MEASURED?

Definition of “importance” and “relevance” is different for different explainers.

*toy examples

Our contribution

Grounding the definition of importance using notions of cause and effect

1. Formulate a **robust metric** (using necessity and sufficiency) which is defined by the ideas of cause and effect. (causality) **to quantify importance**.

2. Unify and **evaluate the robustness of different feature importance ranking algorithms using the concept of necessity and sufficiency.**

Necessity and Sufficiency
Philosophical and casual concepts for cause and effect

Necessity and sufficiency are concepts that have been extensively explored in philosophy, and causal interpretations.

Necessary cause**:**

If **the cause** is **FALSE; the effect** must be **FALSE**

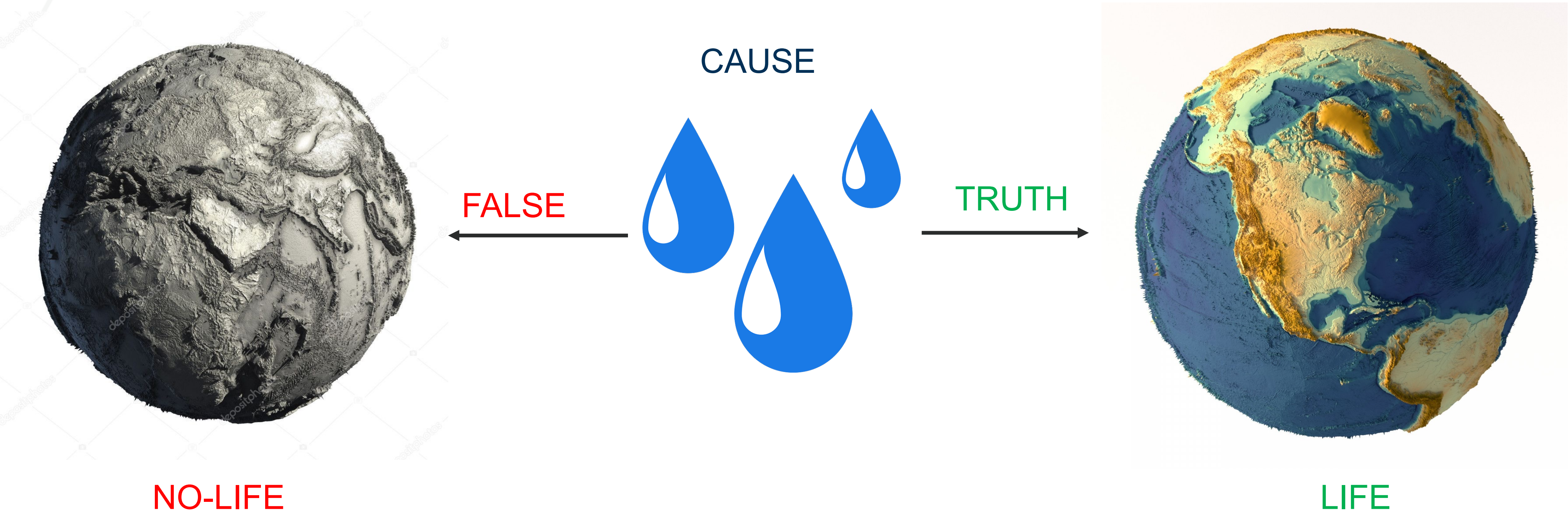
Sufficient cause**:**

If **the cause** is **TRUE; the effect** must be **TRUE**

Swartz, N., 1997, The concepts of necessary conditions and sufficient conditions: Department of Philosophy Simon Fraser University.

Necessary Cause:
If **the cause** is **FALSE; the effect** must be **FALSE, too.**

Water is NECESSARY for life.



Necessary Cause:
If **the cause** is **FALSE; the effect** must be **FALSE, too.**

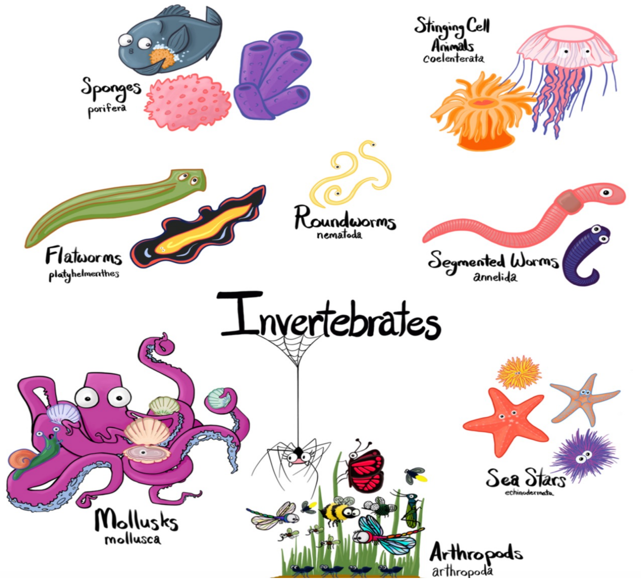
Water is NECESSARY for life.

But it is not sufficient

Sufficient Cause

If **the cause** is **TRUE; the effect** must always be **TRUE.**

Fur on body is SUFFICIENT to be a MAMMAL



NOT MAMMALS

FALSE ←

FUR

→ TRUE

MAMMALS

Sufficient Cause
If **the cause** is **TRUE; the effect** must always be **TRUE.**

Fur on body is SUFFICIENT to be a MAMMAL

STILL MAMMALS

FALSE

FUR

TRUE

MAMMALS

But it is not necessary

Calculating Necessity Score

For an ML classifier the features (or attributes) are the cause, and the outcome is the effect.

Necessity is calculated by making the concerned cause FALSE and checking if effect is FALSE or not.

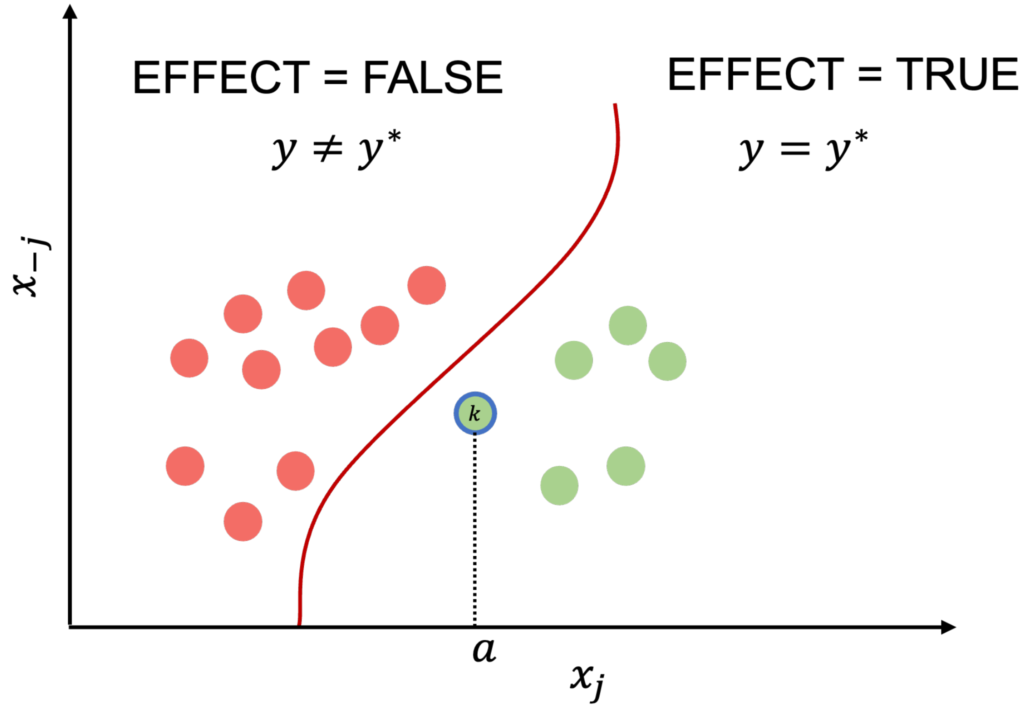
Here: Cause: $x_j = \alpha$ and Effect: y

Initial conditions: Cause: TRUE and Effect: TRUE

$$x_j = \alpha, and\ y = y^*$$

Target conditions: Effect: FALSE when Cause: FALSE

$$y \neq y^* \text{when } x_j \neq \alpha$$



$(b) A\ binary\ classifier\ is\ fit\ on\ this\ datapoint\ for\ outcomes\ (y = y^*\ \&\ y \neq y^*)$

Calculating Necessity Score

For an ML classifier the features (or attributes) are the cause, and the outcome is the effect.

Necessity is calculated by making the concerned cause FALSE and checking if effect is FALSE or not.

Here: Cause: $x_j = \alpha$ and Effect: y

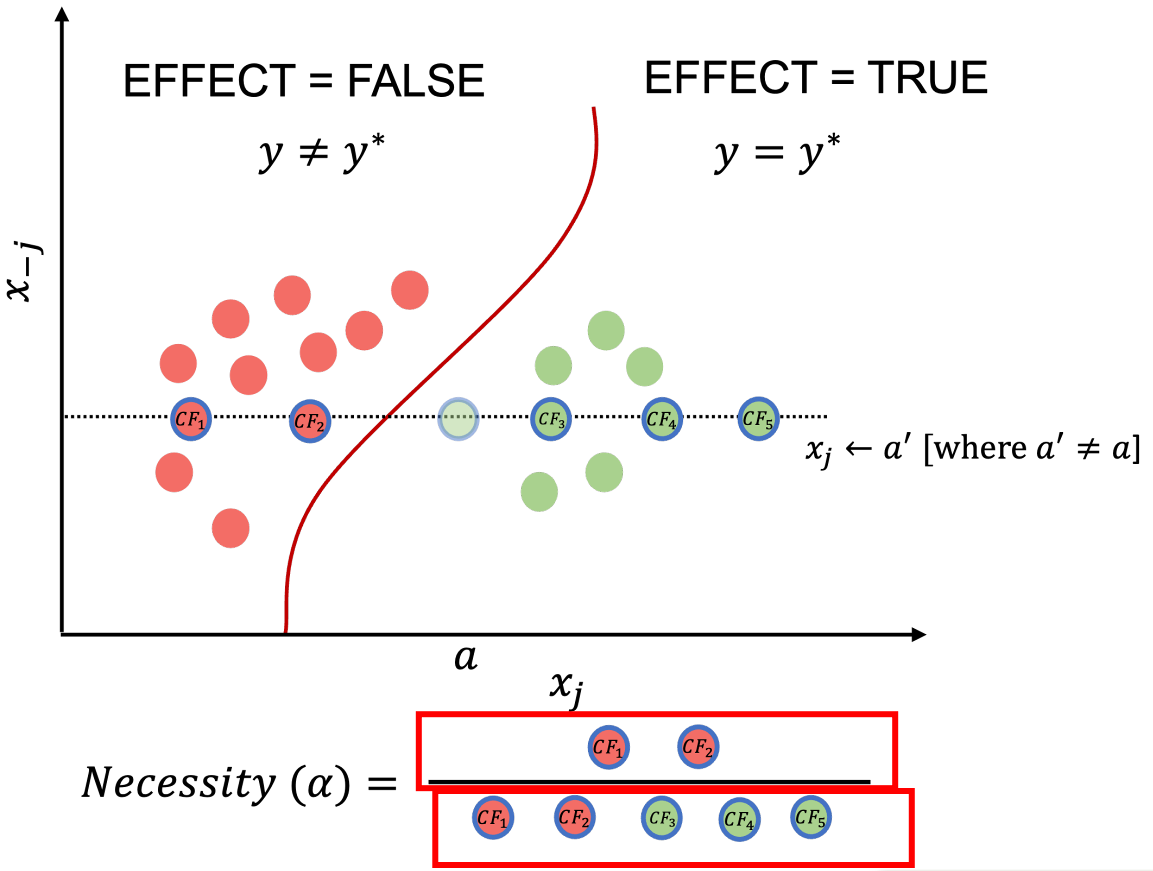
Initial conditions: Cause: TRUE and Effect: TRUE

$$x_j = \alpha, and\ y = y^*$$

Target conditions: Effect: FALSE when Cause: FALSE

$$y \neq y^* \text{when } \mathrm{x_j} \neq \alpha$$

$$\text{Necessity} = \frac{\boxed{\sum^N \sum_i^n \mathbf{1}\left(CF_i(N) \mid x_j \neq a, y \neq y^*\right)}}{\boxed{\mathrm{n} * N}}$$



OLIVES @GeorgiaTech

Georgia Tech

Calculating Sufficiency Score

For an ML classifier the features (or attributes) are the cause, and the outcome is the effect.

Sufficiency is calculated by making a FALSE cause TRUE and checking if effect becomes TRUE or not.

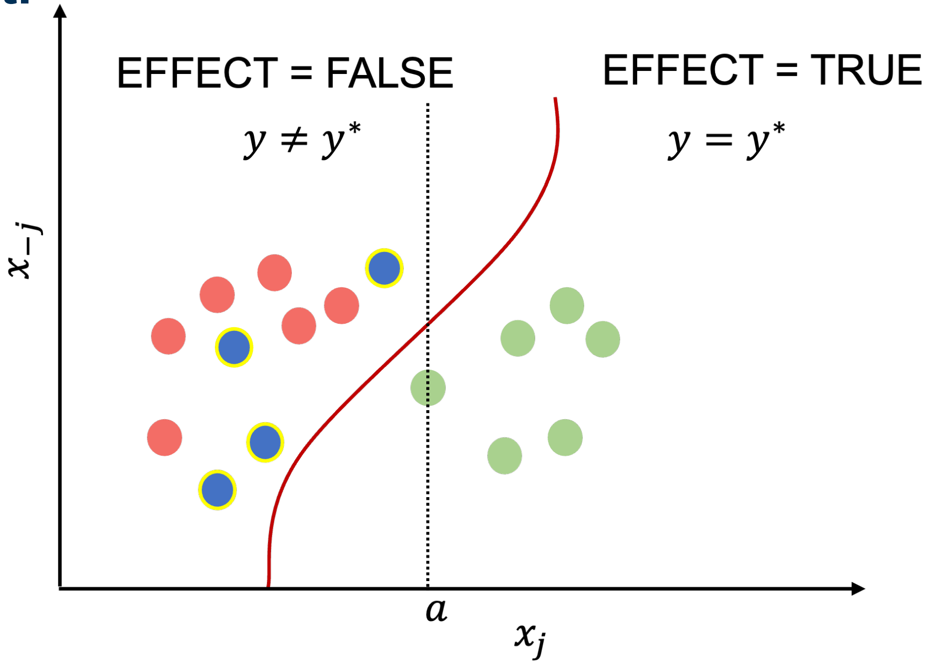
Here: Cause: $x_j = \alpha$ and Effect: y

Initial conditions: Cause: FALSE and Effect: FALSE

$$y \neq y^* \text{when } \mathrm{x_j} \neq \alpha$$

Target conditions: Effect: TRUE when Cause: TRUE

$$x_j = \alpha, and\ y = y^*$$



$(b) A\ binary\ classifier\ is\ fit\ on\ this\ datapoint\ for\ outcomes\ (y = y^*\ \&\ y \neq y^*)$

Calculating Sufficiency Score

For an ML classifier the features (or attributes) are the cause, and the outcome is the effect.

Sufficiency is calculated by making a FALSE cause TRUE and checking if effect becomes TRUE or not.

Here: Cause: $x_j = \alpha$ and Effect: y

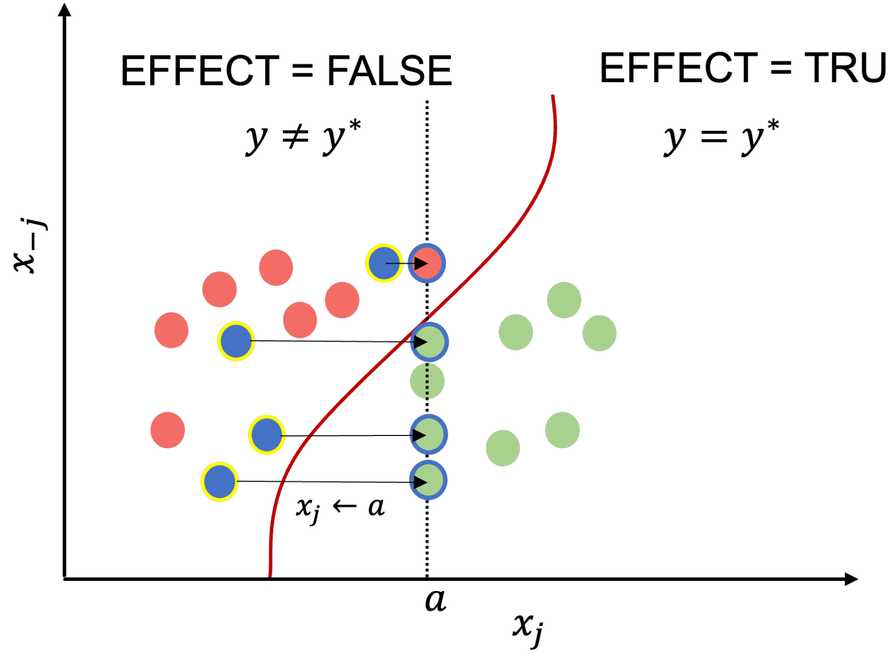
Initial conditions: Cause: FALSE and Effect: FALSE

$$y \neq y^* \text{when } \mathrm{x_j} \neq \alpha$$

Target conditions: Effect: TRUE when Cause: TRUE

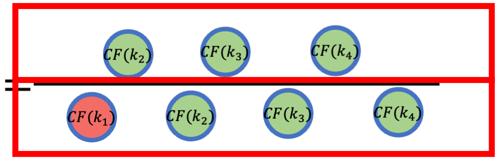
$$x_j = \alpha, and\ y = y^*$$

$$\text{Sufficiency} = \frac{\sum^R \sum^K \mathbf{1}\left(CF(k) \mid x_j \leftarrow a, y = y^*\right)}{K * R}$$



$(b) A\ binary\ classifier\ is\ fit\ on\ this\ datapoint\ for\ outcomes\ (y = y^*\ \&\ y \neq y^*)$

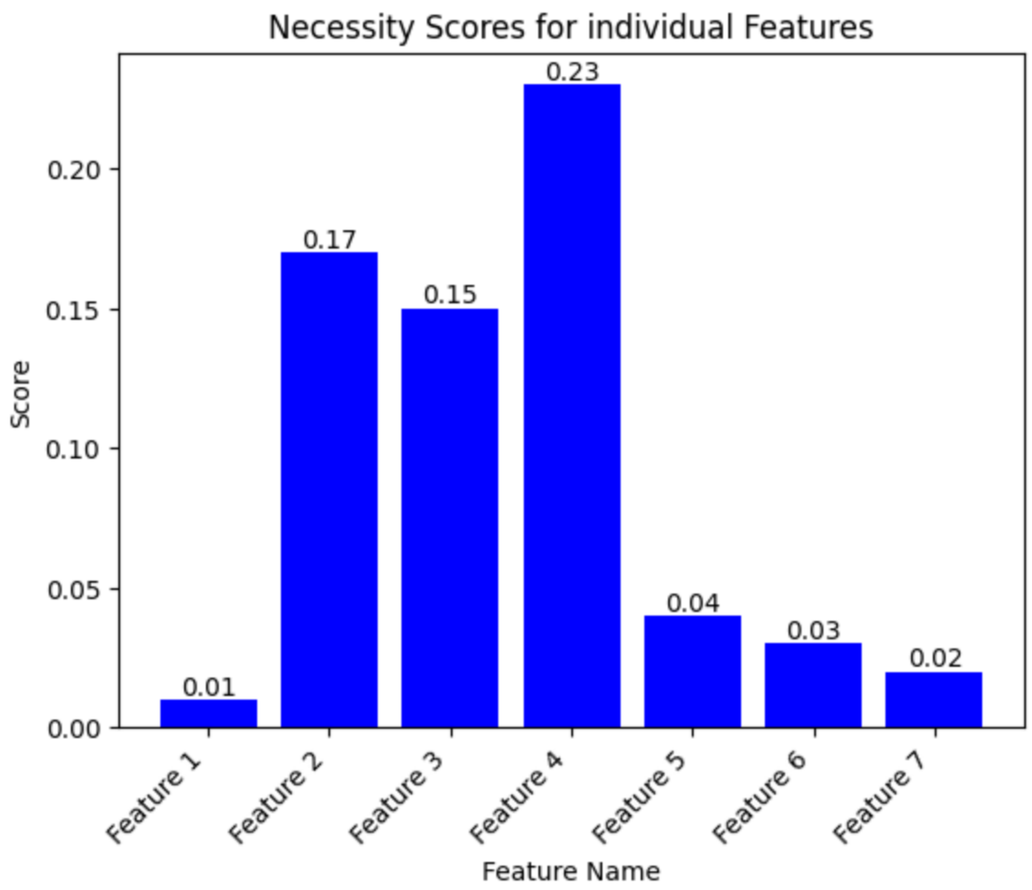
$Sufficiency\ (\beta) =$

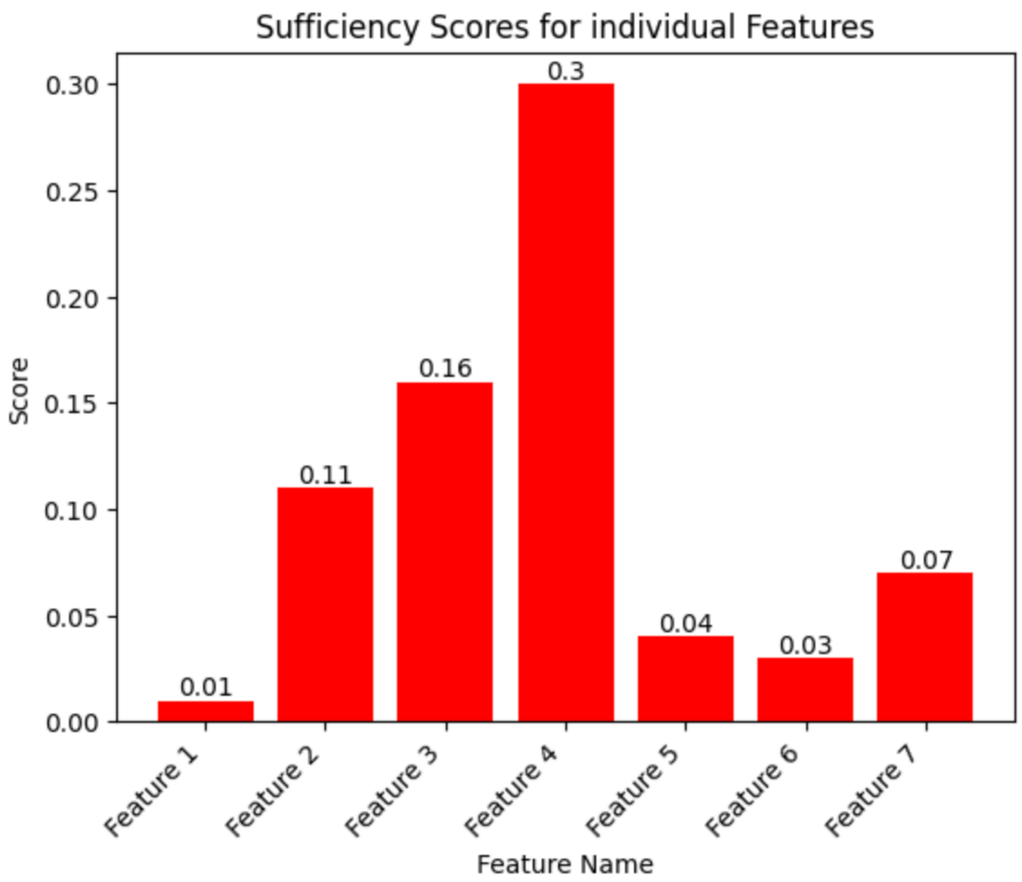


Necessity and Sufficiency Scores

Each feature of a dataset can be assigned its individual necessity and sufficiency scores

The below bar graphs displays the corresponding scores for the top 7 features in the DHI dataset





*Actual names not shown for data confidentiality

Analysis of LIME and SHAP explanations

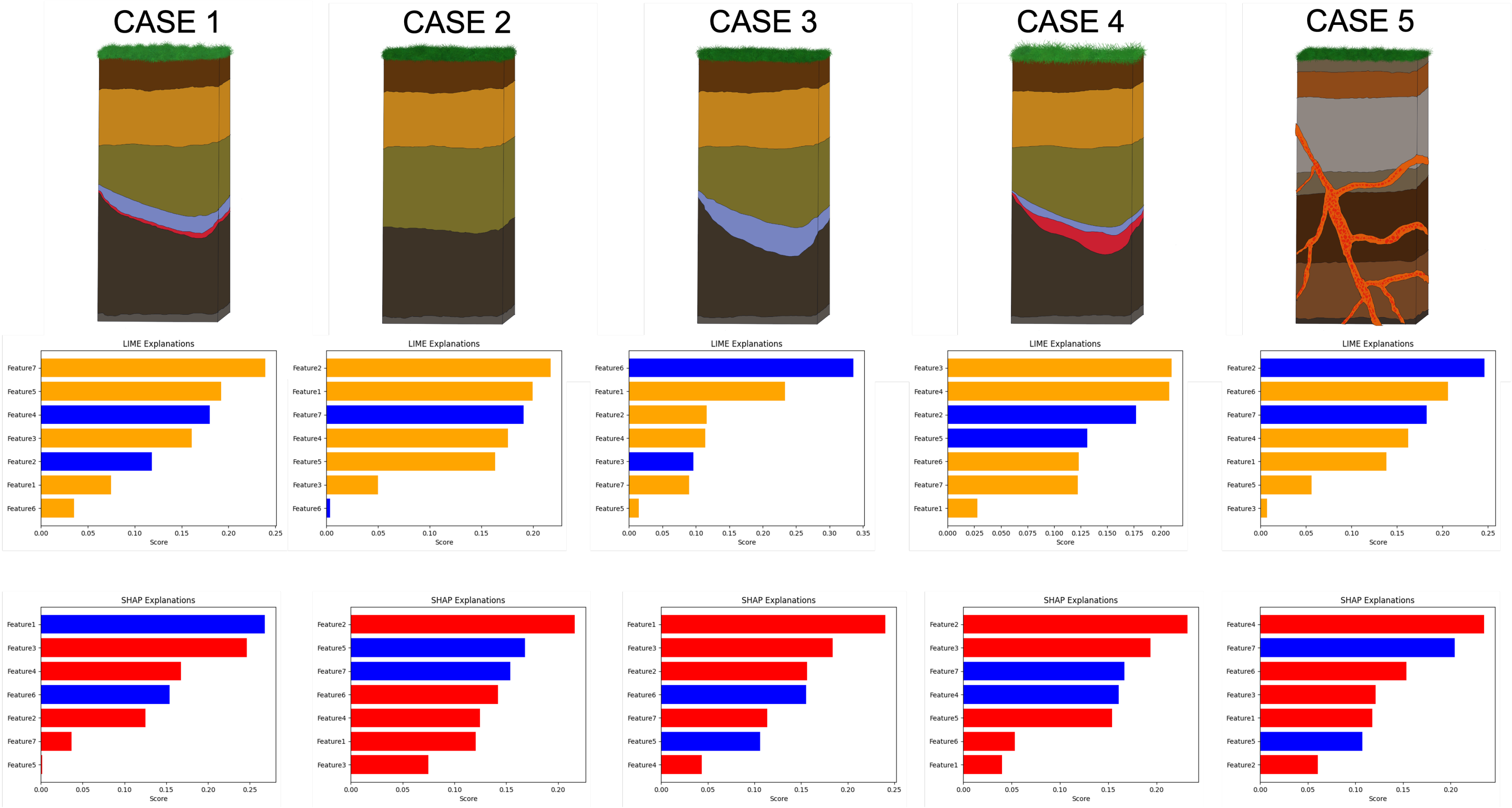
Towards verifying the robustness of the feature importance rankings by these XAI methods

Different datapoints (cases) generate different LIME and SHAP explanation

Individual Prospects

LIME Explanations

SHAP Explanations



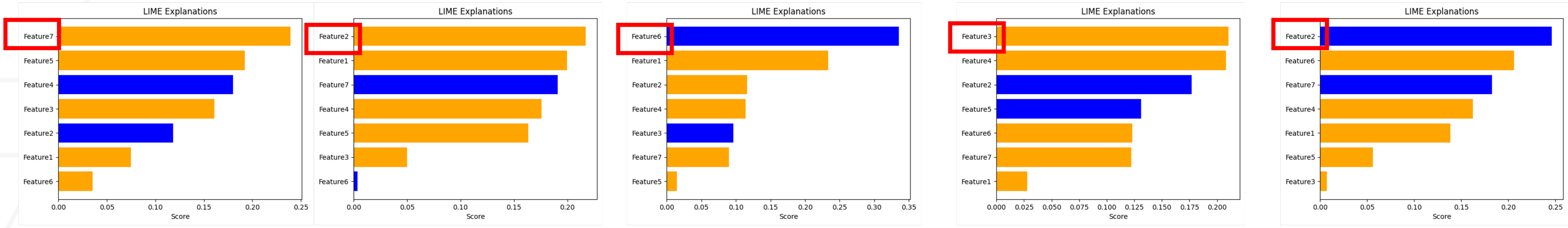
*toy examples

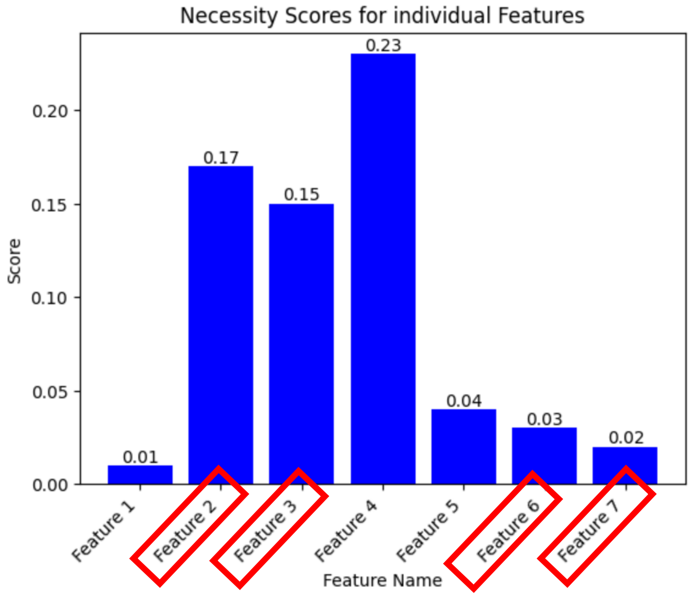
Analysis of LIME and SHAP explanations

Towards verifying the robustness of the feature importance rankings by these XAI methods

Average the necessity (or sufficiency) score for each feature Ranked #1 for each LIME (or SHAP) explanation.

CASE 1 CASE 2 CASE 3 CASE 4 CASE 5





Rank#1 Necessity score (NS) =

$$\frac{NS_{Feature7} + NS_{Feature2} + NS_{Feature6} + NS_{Feature3} + NS_{Feature2}}{5}$$

(For LIME explanations for DHI Data)

*for experiments it is averaged over all test data points

OLIVES @GeorgiaTech

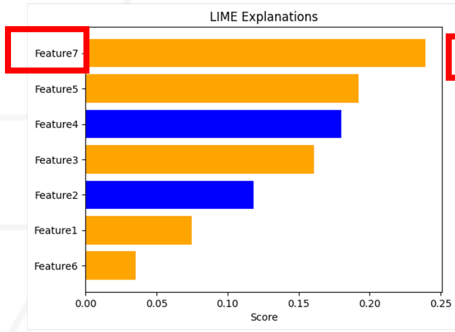
Georgia Tech

Analysis of LIME and SHAP explanations

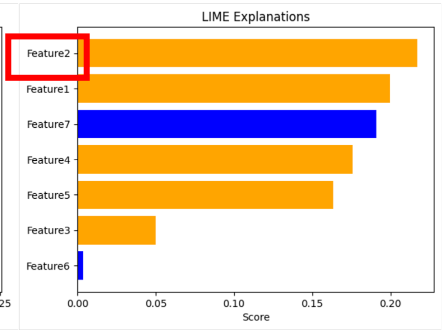
Towards verifying the robustness of the feature importance rankings by these XAI methods

Average the necessity (or sufficiency) score for each feature Ranked #1 for each LIME (or SHAP) explanation.

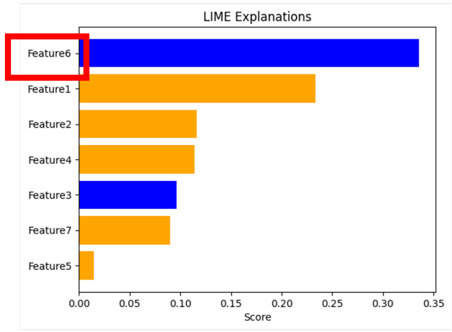
CASE 1



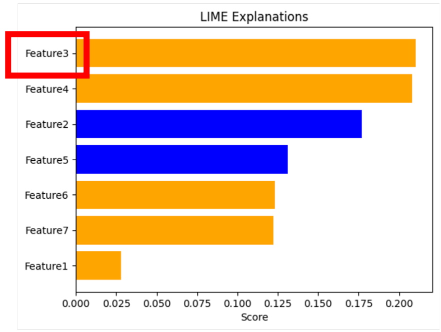
CASE 2



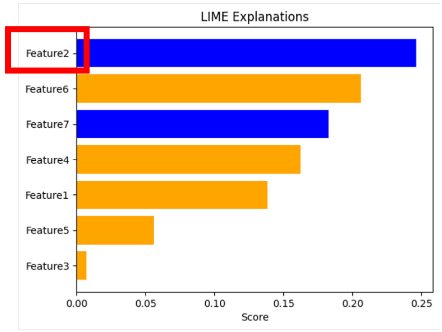
CASE 3

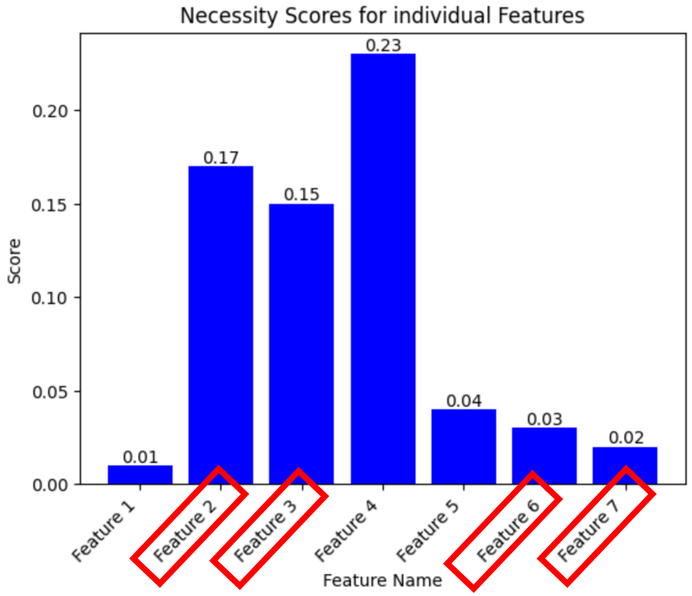


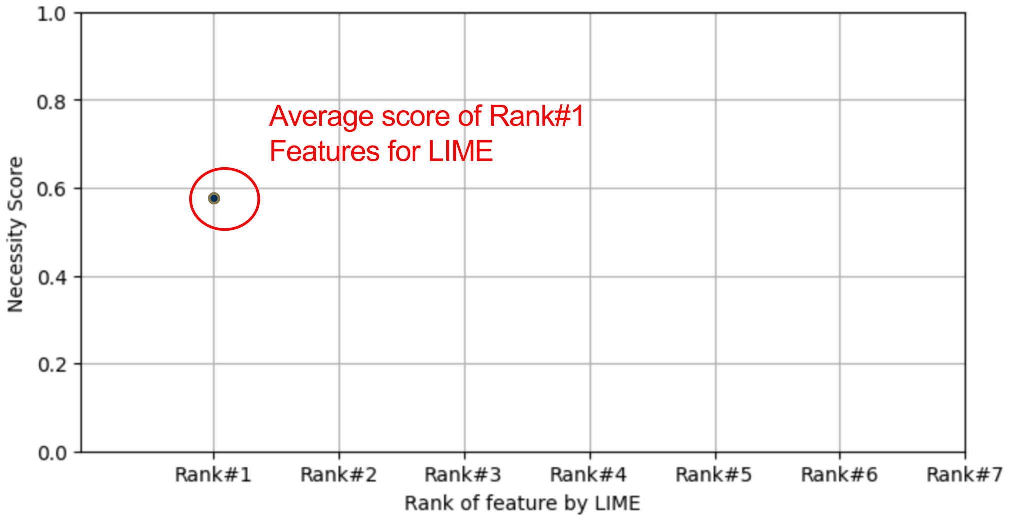
CASE 4



CASE 5







*toy examples

OLIVES @GeorgiaTech

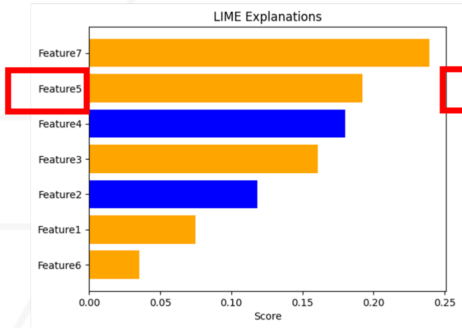
Georgia Tech

Analysis of LIME and SHAP explanations

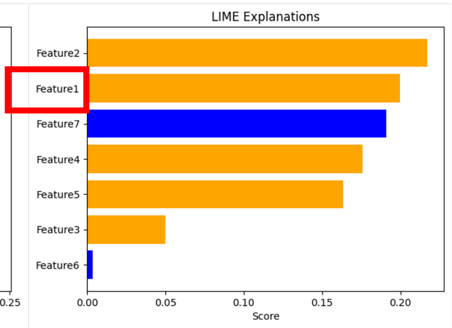
Towards verifying the robustness of the feature importance rankings by these XAI methods

Average the necessity (or sufficiency) score for each feature Ranked #2 for each LIME (or SHAP) explanation.

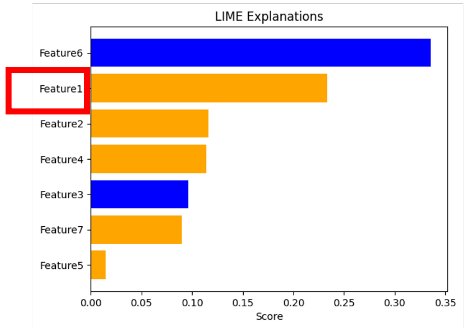
CASE 1



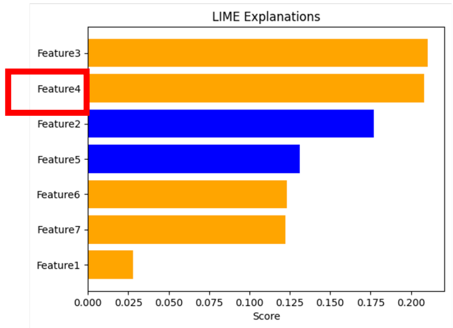
CASE 2



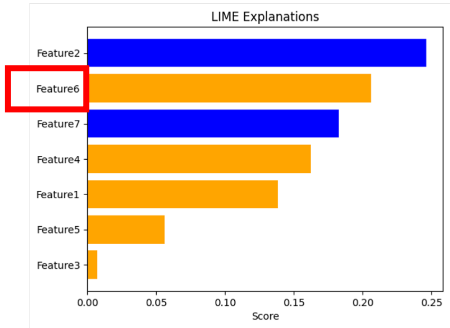
CASE 3

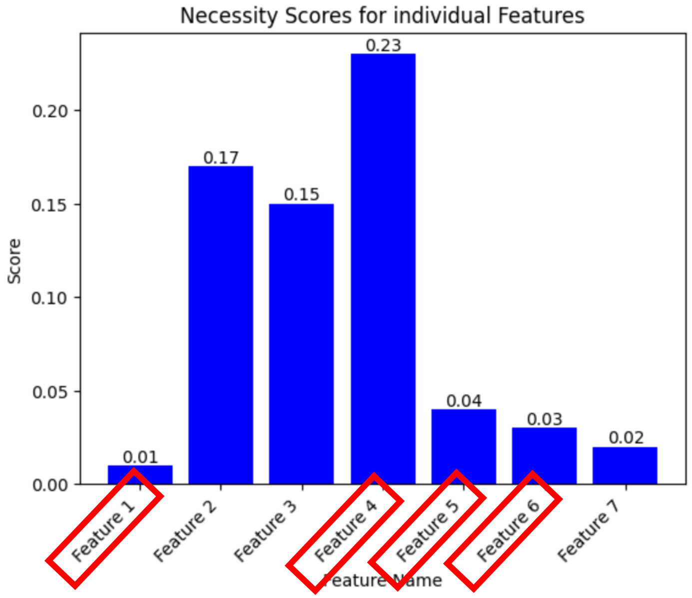


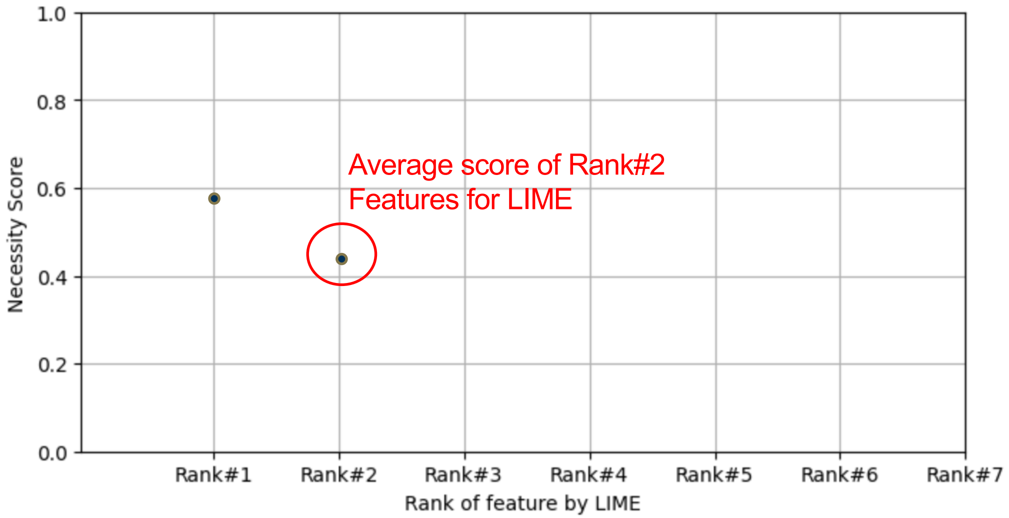
CASE 4



CASE 5







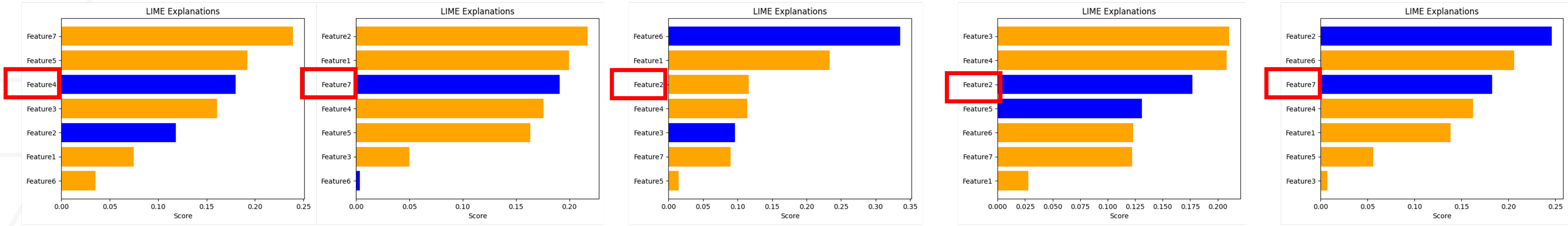
*toy examples

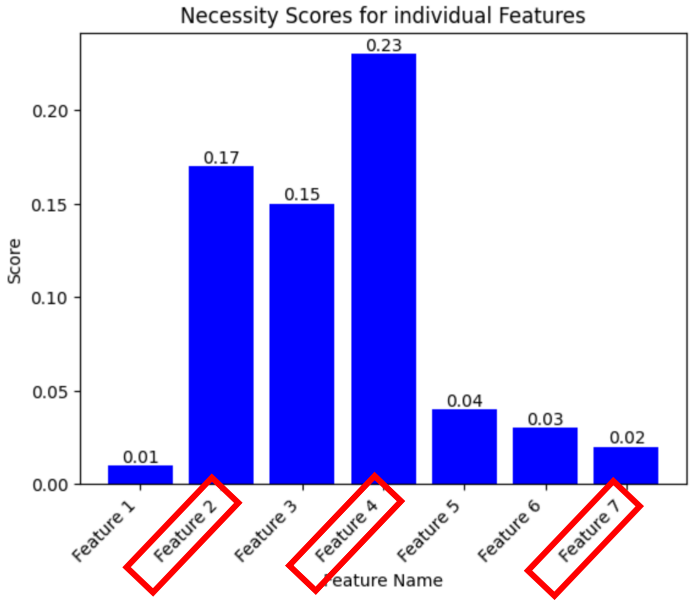
Analysis of LIME and SHAP explanations

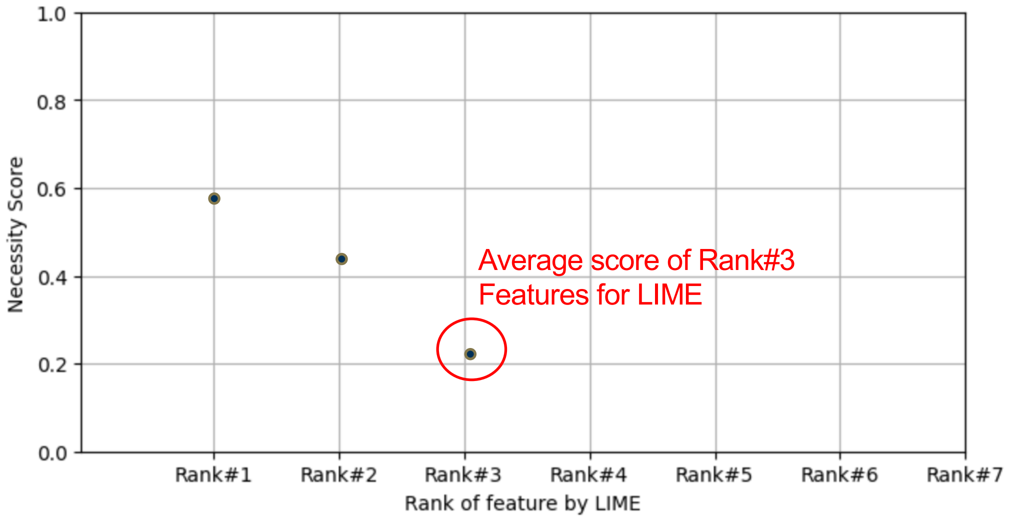
Towards verifying the robustness of the feature importance rankings by these XAI methods

Average the necessity (or sufficiency) score for each feature Ranked #3 for each LIME (or SHAP) explanation.

CASE 1 CASE 2 CASE 3 CASE 4 CASE 5







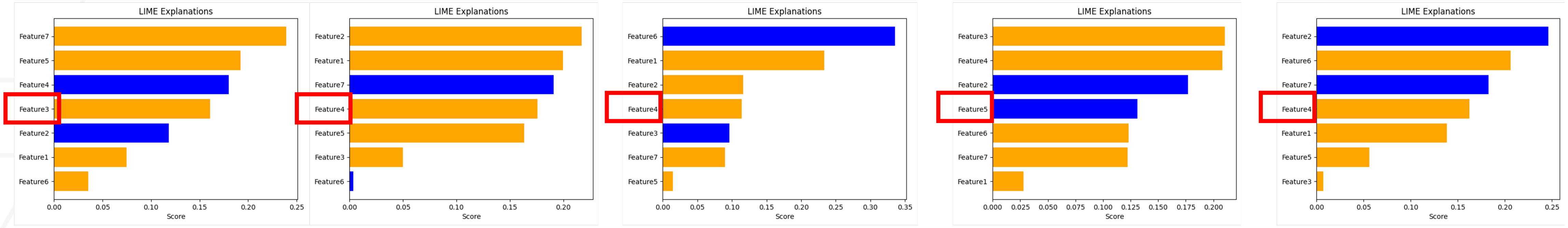
*toy examples

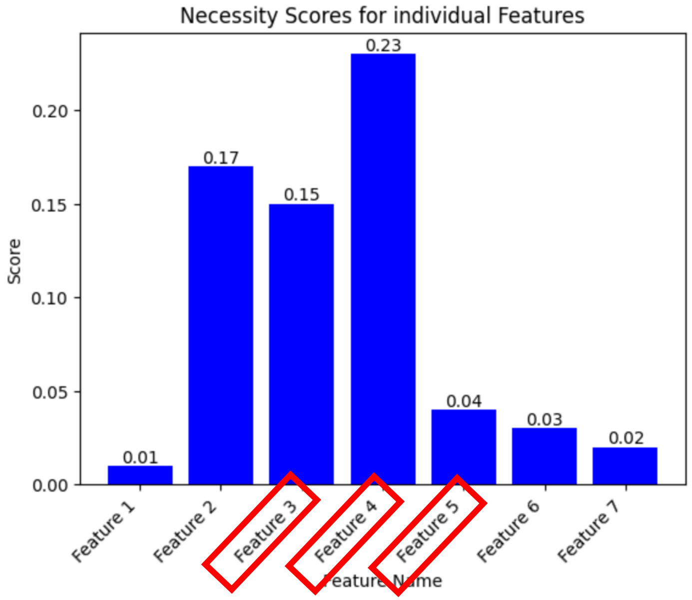
Analysis of LIME and SHAP explanations

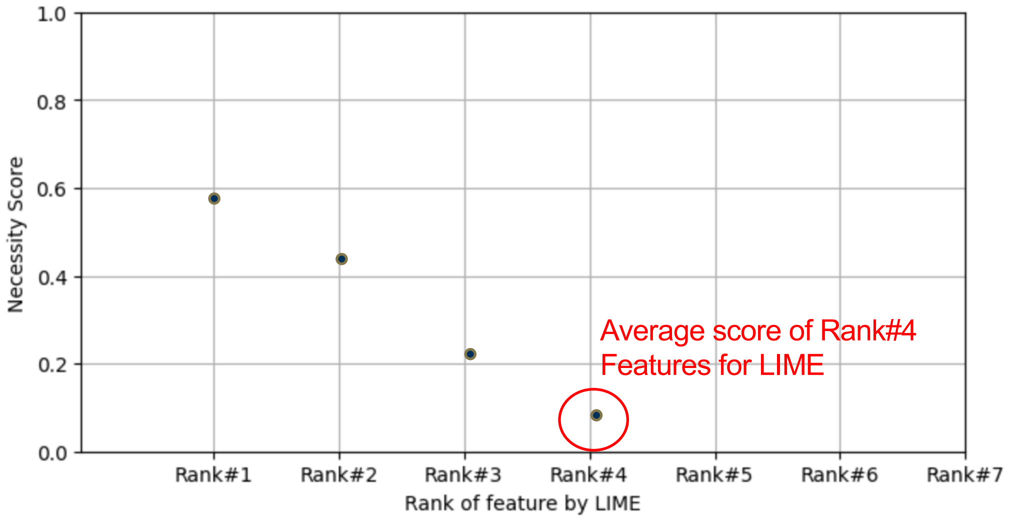
Towards verifying the robustness of the feature importance rankings by these XAI methods

Average the necessity (or sufficiency) score for each feature Ranked #4 for each LIME (or SHAP) explanation.

CASE 1 CASE 2 CASE 3 CASE 4 CASE 5







*toy examples

OLIVES @GeorgiaTech

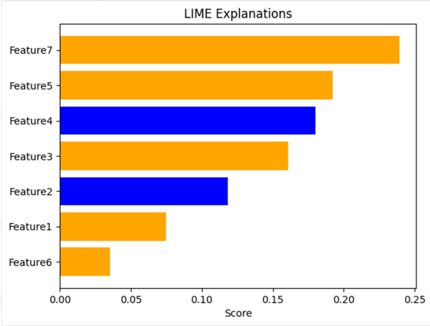
Georgia Tech

Analysis of LIME and SHAP explanations

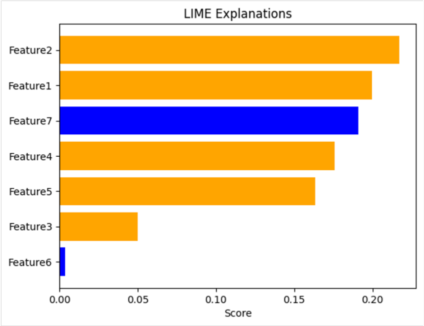
Towards verifying the robustness of the feature importance rankings by these XAI methods

Ideal SCENRIO: The scores should be monotonously decreasing with rank

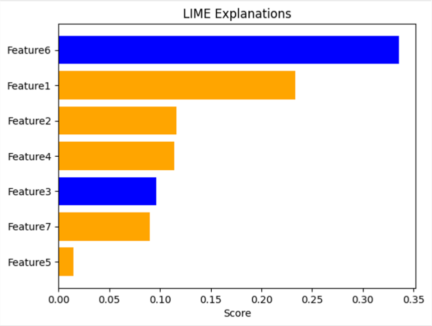
CASE 1



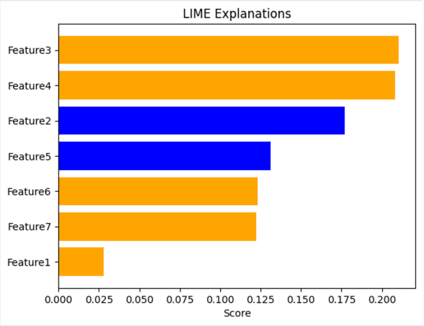
CASE 2



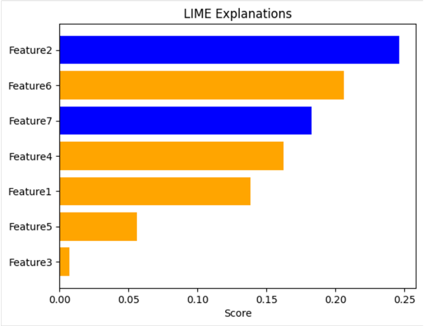
CASE 3

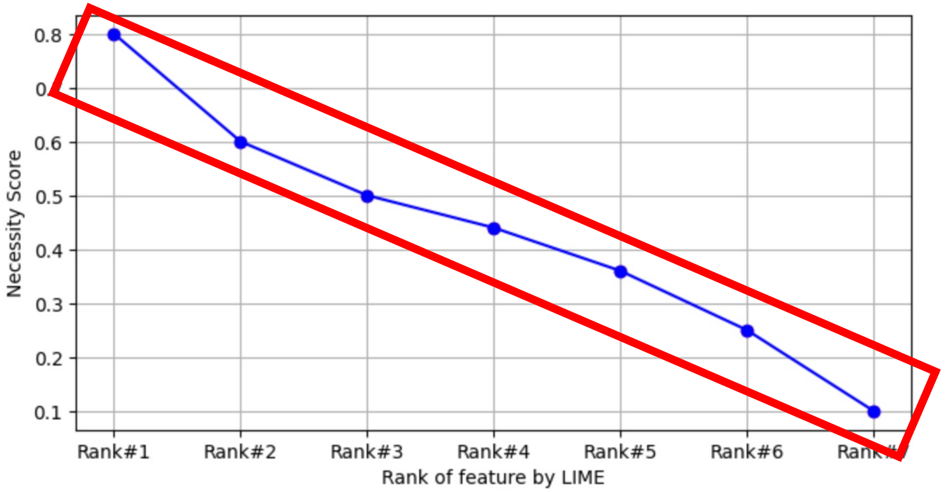


CASE 4



CASE 5





For an **explanation** to be **robust**:

An **important feature** should be **proportionately necessary and sufficient**.

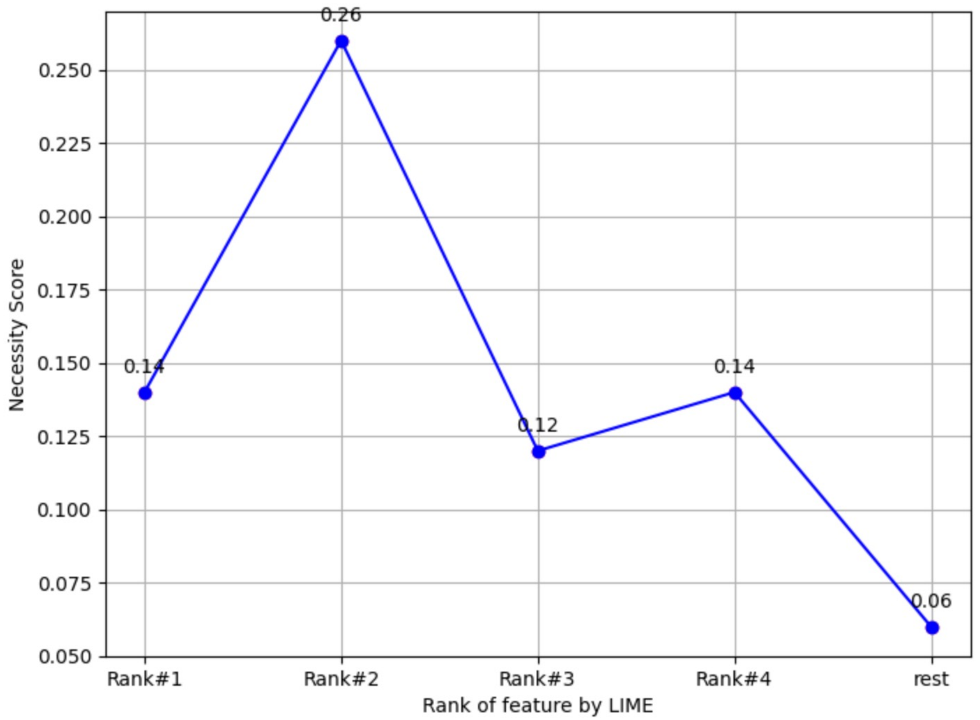
*toy examples

OLIVES @GeorgiaTech

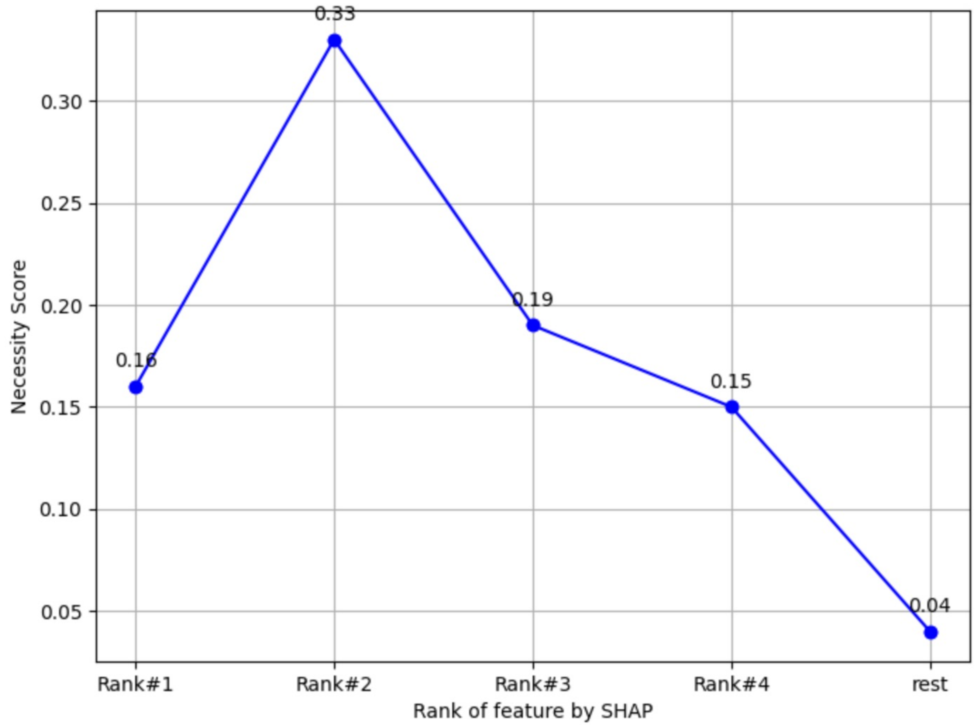
Georgia Tech

Analysis of LIME and SHAP explanations
Towards verifying the robustness of the feature importance rankings by these XAI methods

The LIME and SHAP explanations for DHI data isn't perfectly robust



LIME – Necessity Evaluation

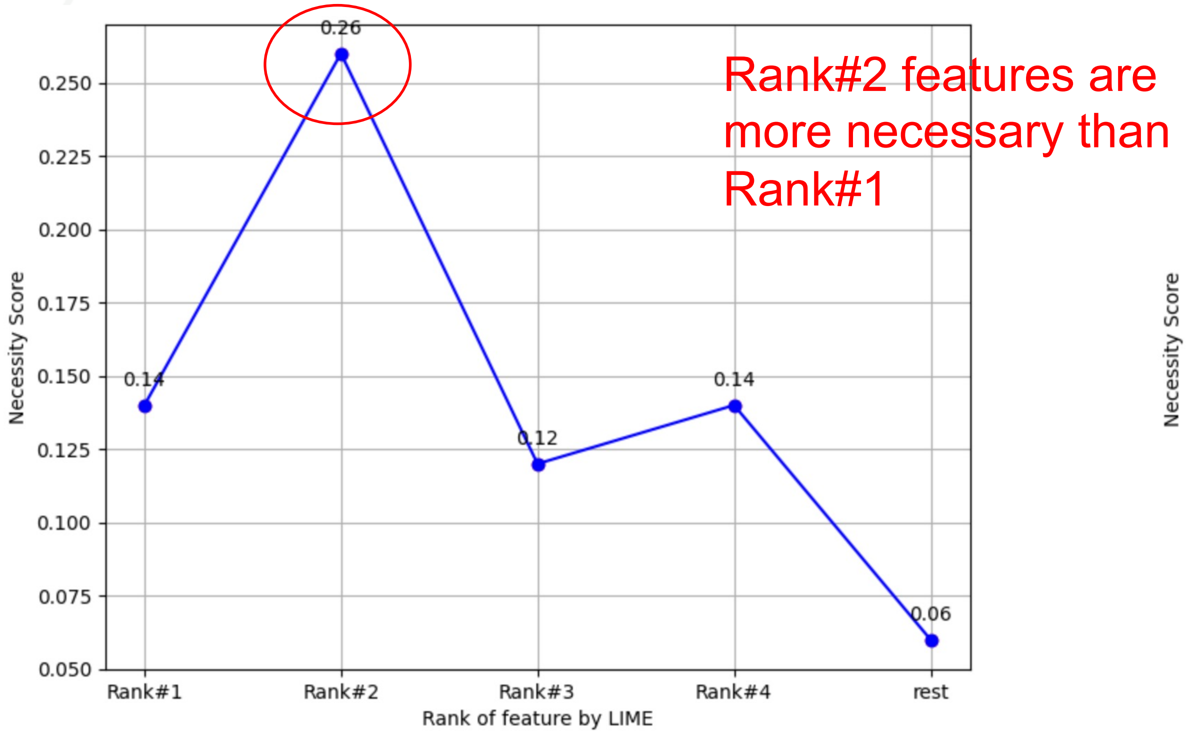


SHAP – Necessity Evaluation

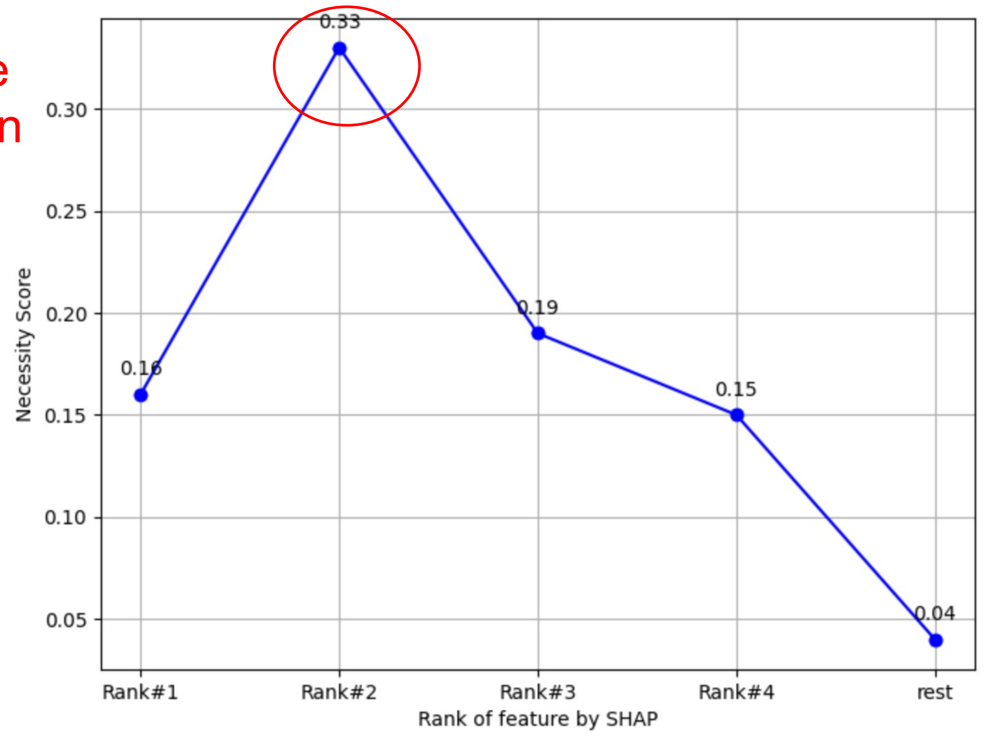
Analysis of LIME and SHAP explanations

Towards verifying the robustness of the feature importance rankings by these XAI methods

The LIME and SHAP explanations for DHI data isn’t perfectly robust to necessity evaluation

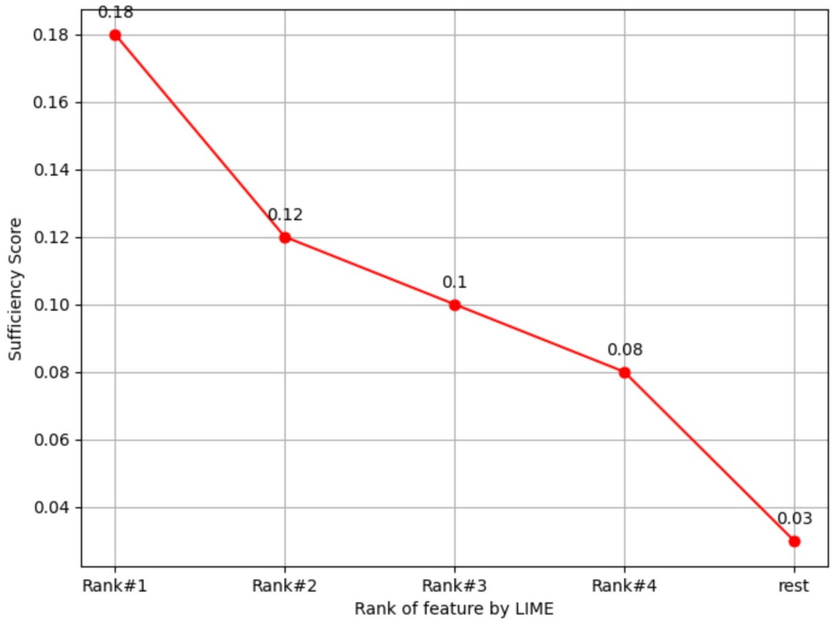


LIME – Necessity Evaluation

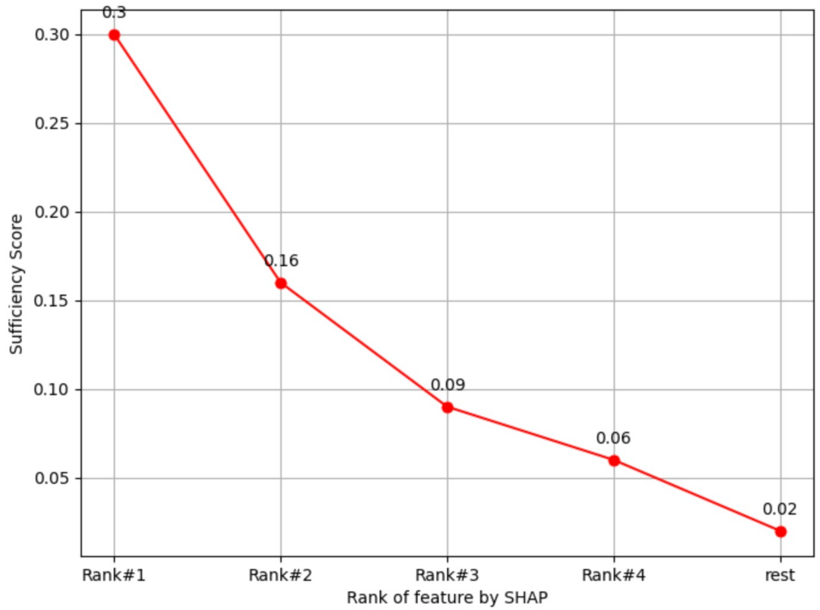


SHAP – Necessity Evaluation

The LIME and SHAP explanations for DHI data is more robust to sufficiency evaluation



LIME – Sufficiency Evaluation



SHAP – Sufficiency Evaluation

For DHI DATA: The importance score assigned by LIME and SHAP explanations to a feature correspond to how sufficient it is for the outcome prediction.

Conclusions

To properly analyze the behavior of an ML model the employment of several explanation methods backed by theoretical concepts are useful.

- We provide a causally defined metric to calculate the impact of each individual features in a dataset for a model's decision.
- We provide a proper evaluation process for the robustness of different local explanation techniques
- Our study grounds the definition of importance as indicated by the local XAI modules for each different scenarios.

Relevant Publications and Codes

P. Chowdhury, M. Prabhushankar, and G. AlRegib, “Explaining Explainers: Necessity and Sufficiency in Tabular Data”, *NeurIPS 2023 Workshop: Table Representation Learning,* submitted on Oct. 4, 2023.

P. Chowdhury, A. Mustafa, M. Prabhushankar and G. AlRegib, "Counterfactual Uncertainty for High Dimensional Structured Datasets," at International Meeting for Applied Geoscience & Energy (IMAGE) 2023, Houston, TX, Aug. 28-Sept. 1, 2023.

For more OLIVES content, please visit:

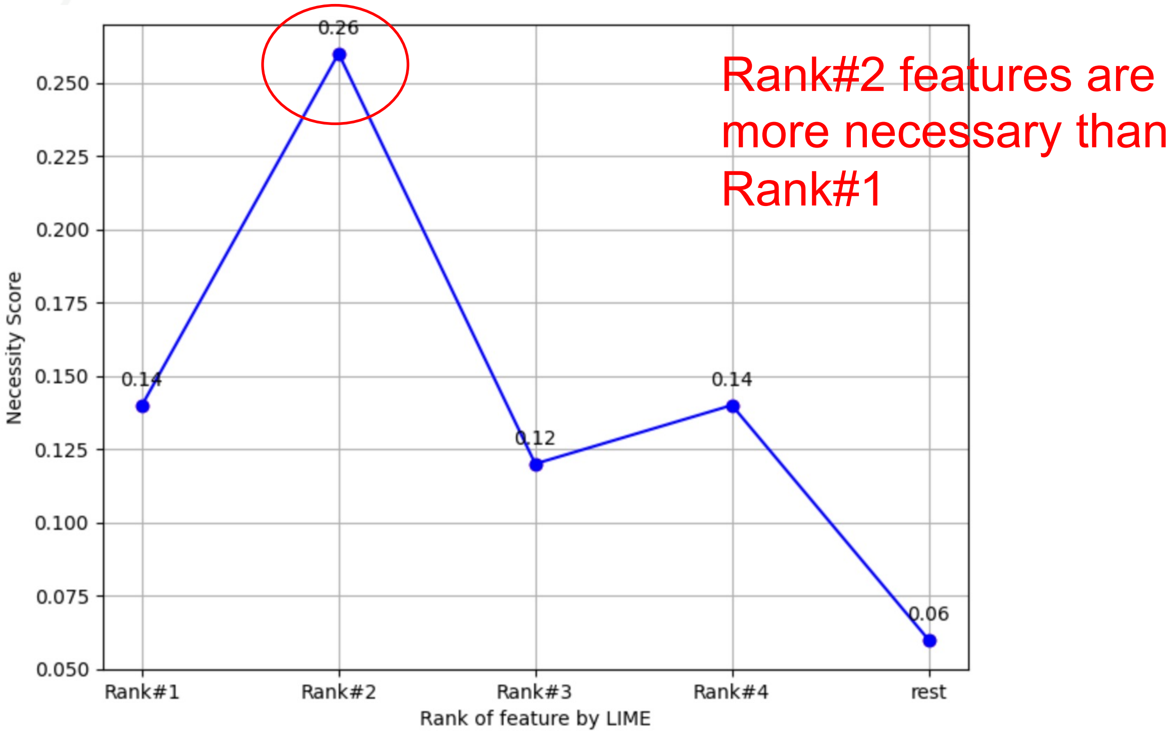
GitHub

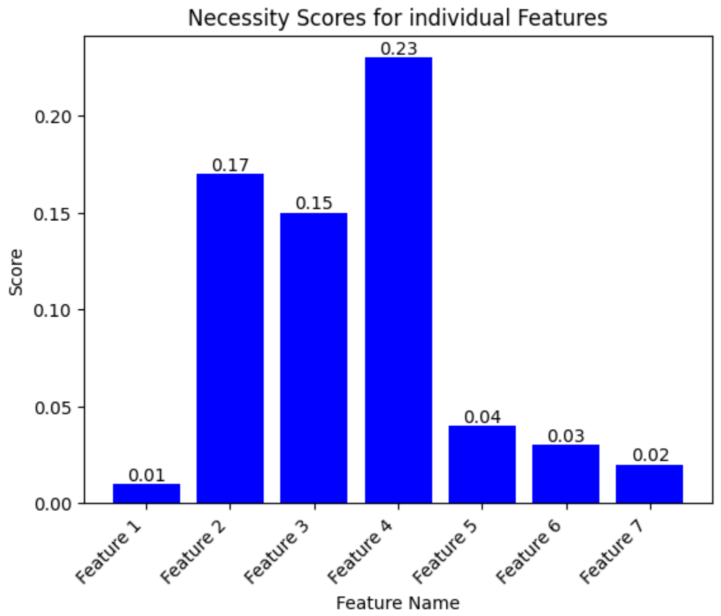
SCAN ME

Publications

SCAN ME

The LIME and SHAP explanations for DHI data isn't perfectly robust to necessity evaluation



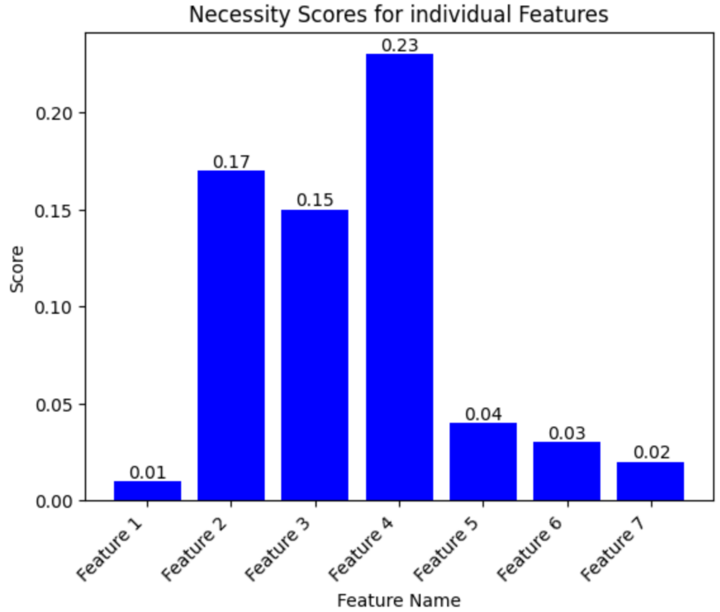


Feature Name	Rank #1 Occurrence	Rank #2 Occurrence	Rank #3 Occurrence
Feature 3	6	38	11
Feature 4	48	13	16

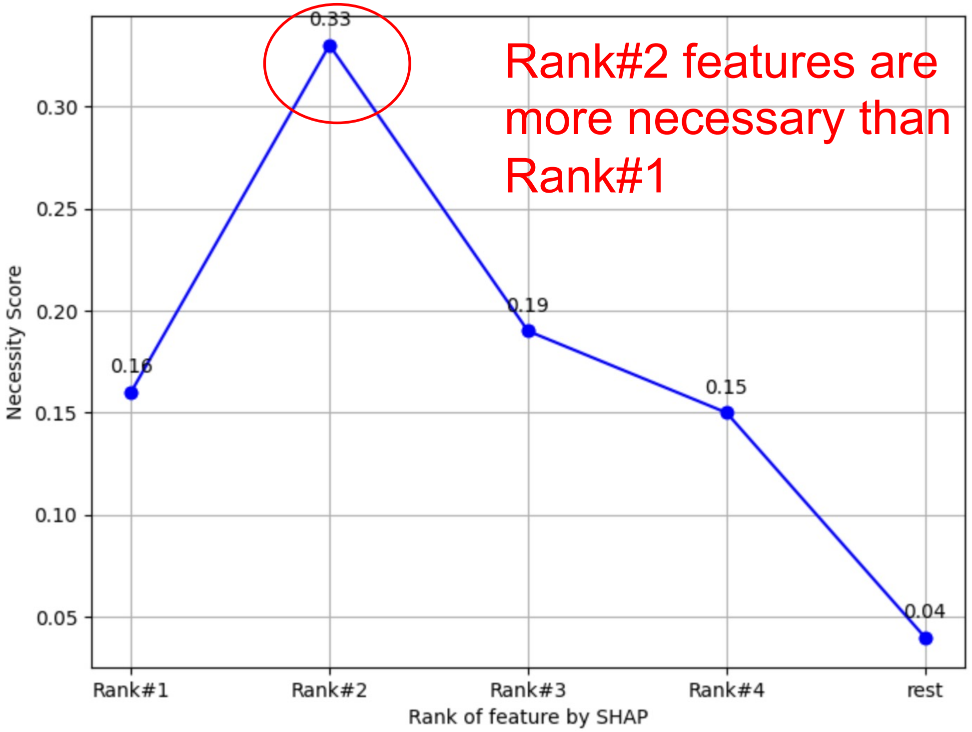
Analysis of LIME and SHAP explanations

Towards verifying the robustness of the feature importance rankings by these XAI methods

The LIME and SHAP explanations for DHI data isn't perfectly robust to necessity evaluation



Feature Name	Rank #1 Occurrence	Rank #2 Occurrence	Rank #3 Occurrence
Feature 3	3	50	22
Feature 4	71	15	2

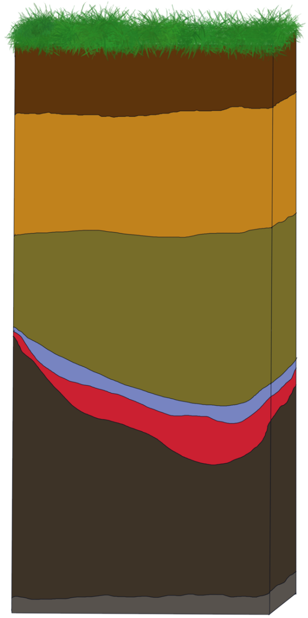


Necessary Cause:
If **the cause** is **FALSE; the effect** must be **FALSE, too.**

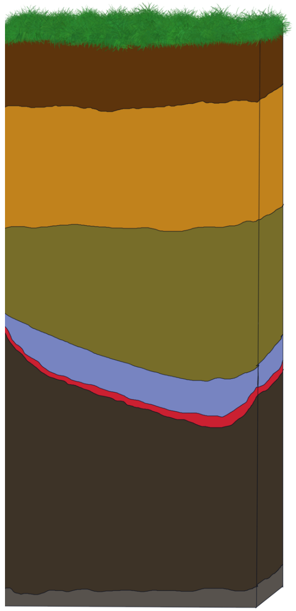
DQ Fluid in the range (0.45 to 0.65) is necessary for a positive prospect outcome

CAUSE = TRUE (DQ Fluid is in range)

DQ Fluid = 0.51



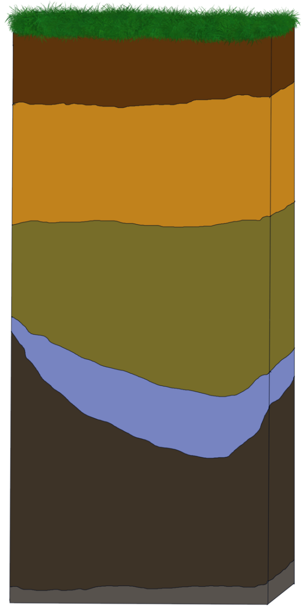
DQ Fluid = 0.63



Outcome: SUCCESS

CAUSE = FALSE (DQ Fluid is out of range)

DQ Fluid = 0.7



DQ Fluid = 0.33

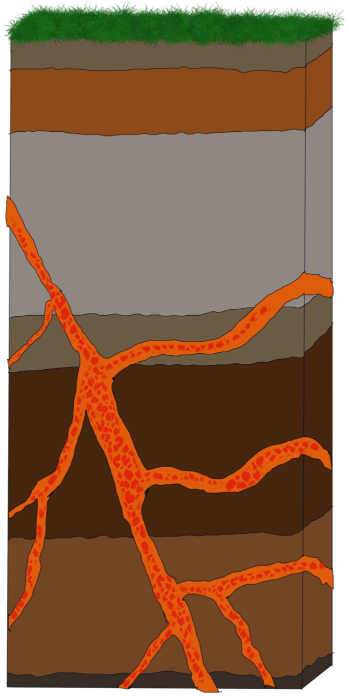
Outcome: FAILURE

*toy examples

Sufficient Cause
If **the cause** is **TRUE; the effect** must always be **TRUE.**

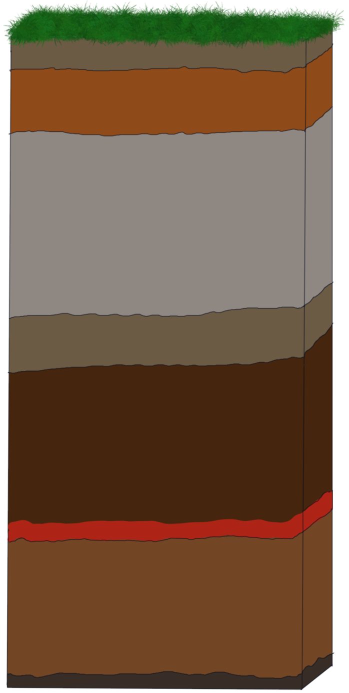
IGN_abs (absence of igneous rock) is a sufficient cause for positive prospect outcome

IGN_abs = 0 (CAUSE = FALSE)



Outcome: FAILURE

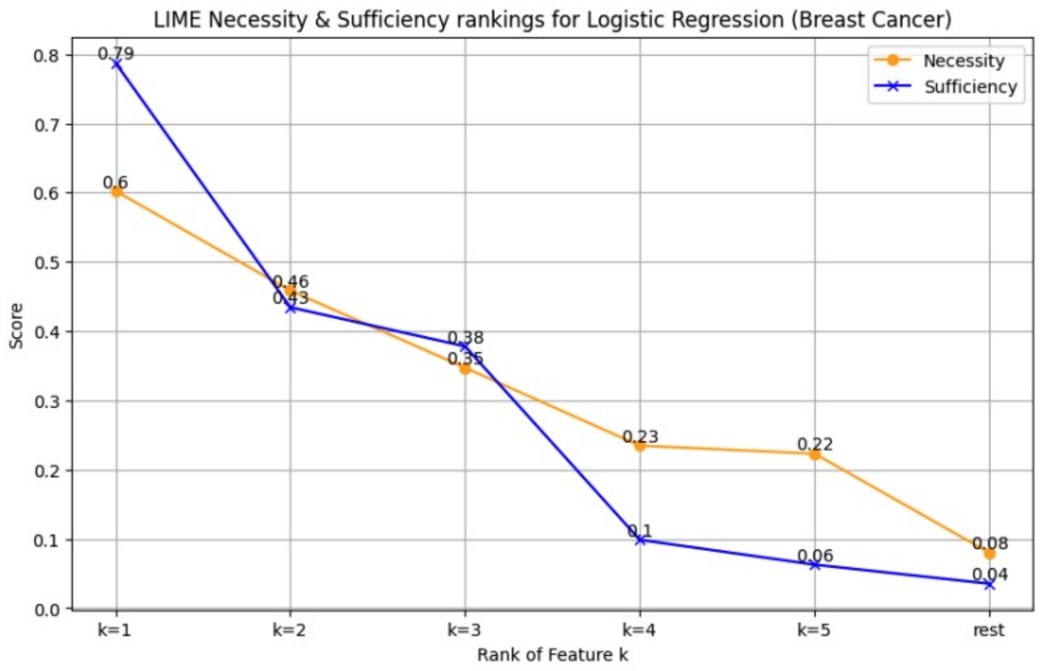
IGN_abs = 1 (CAUSE = TRUE)



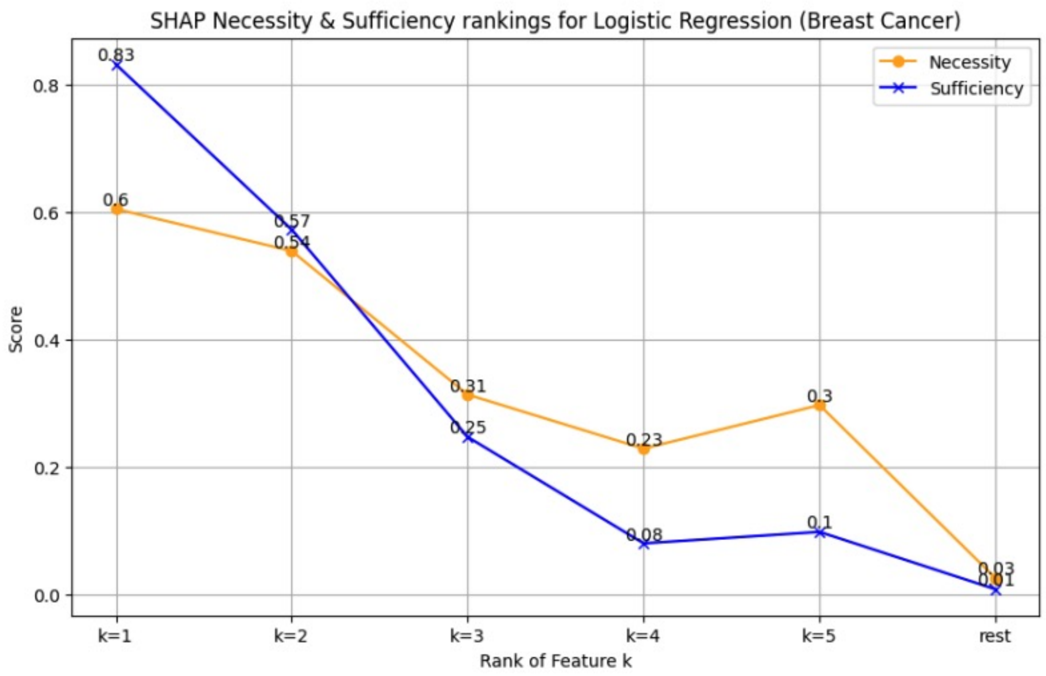
Outcome: SUCCESS

*toy examples

Analysis of LIME and SHAP on Breast Cancer Dataset

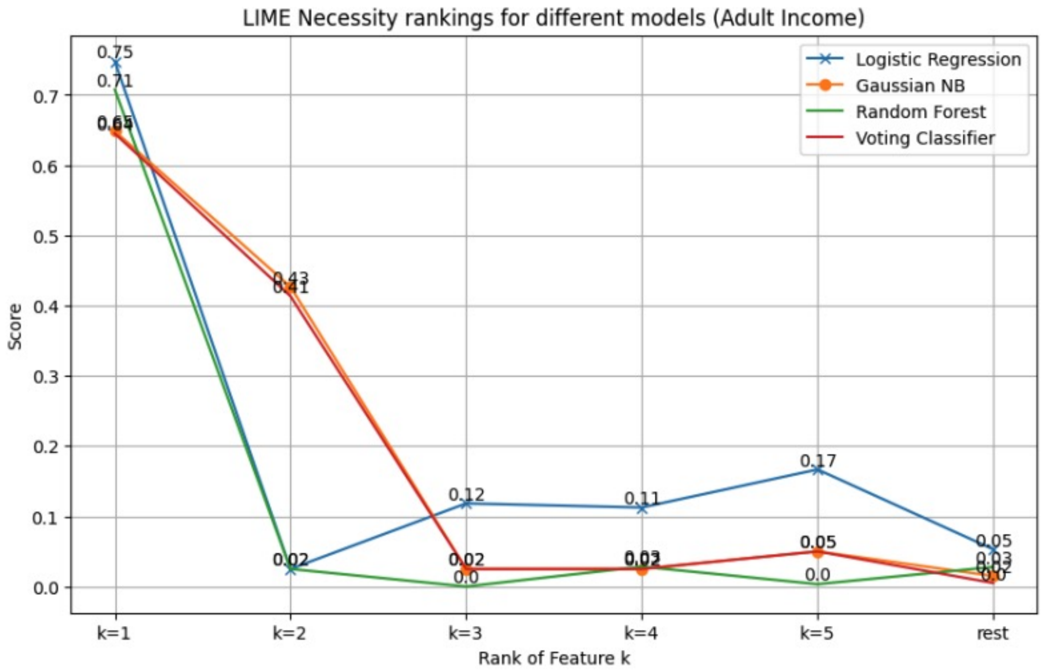


(a) LIME - Necessity & Sufficiency

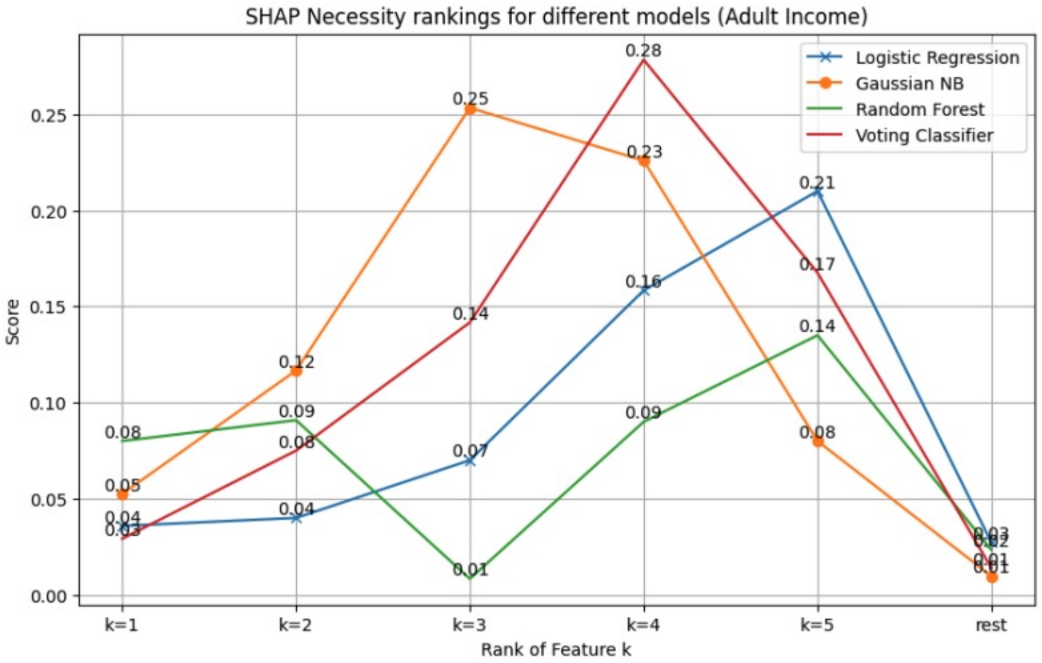


(b) SHAP - Necessity & Sufficiency

Analysis of LIME and SHAP on Adult Income Dataset

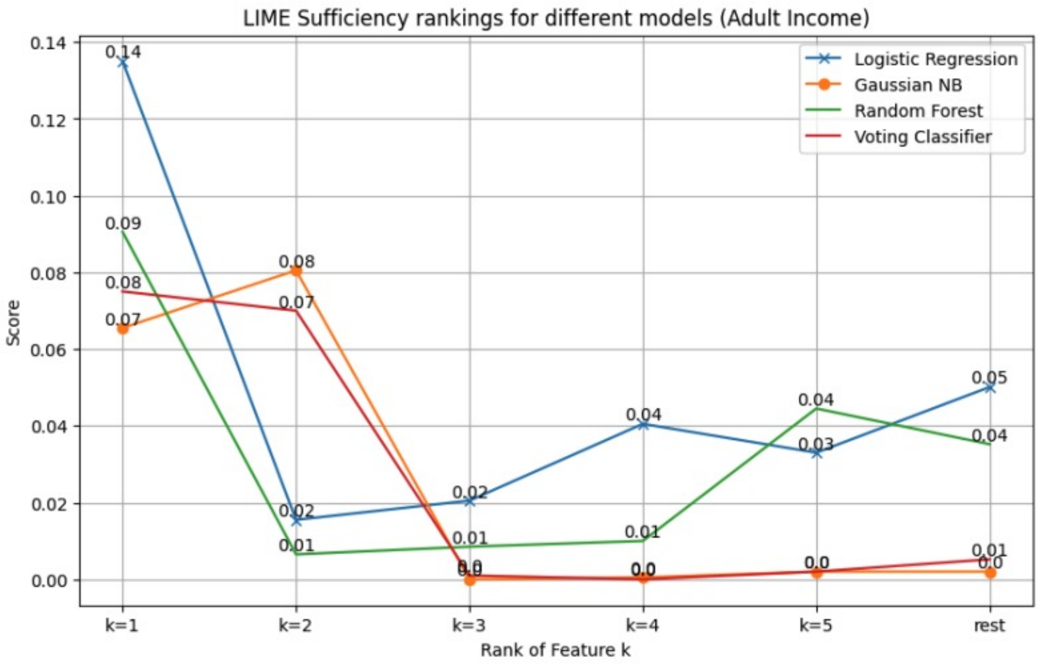


(a) LIME - Necessity

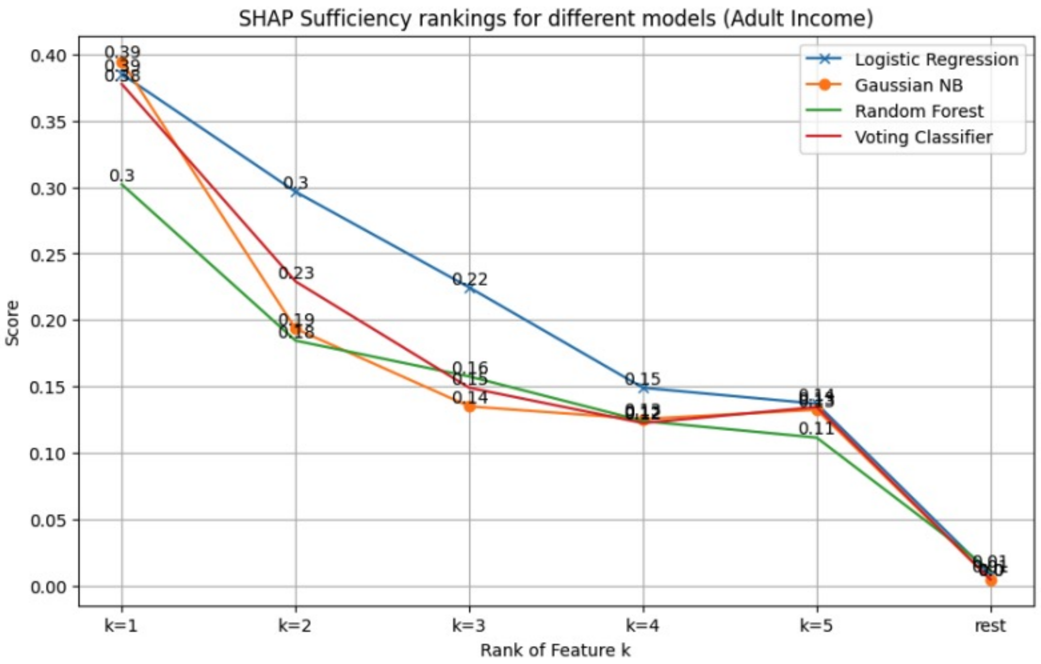


(b) SHAP - Necessity

Analysis of LIME and SHAP on Adult Income Dataset

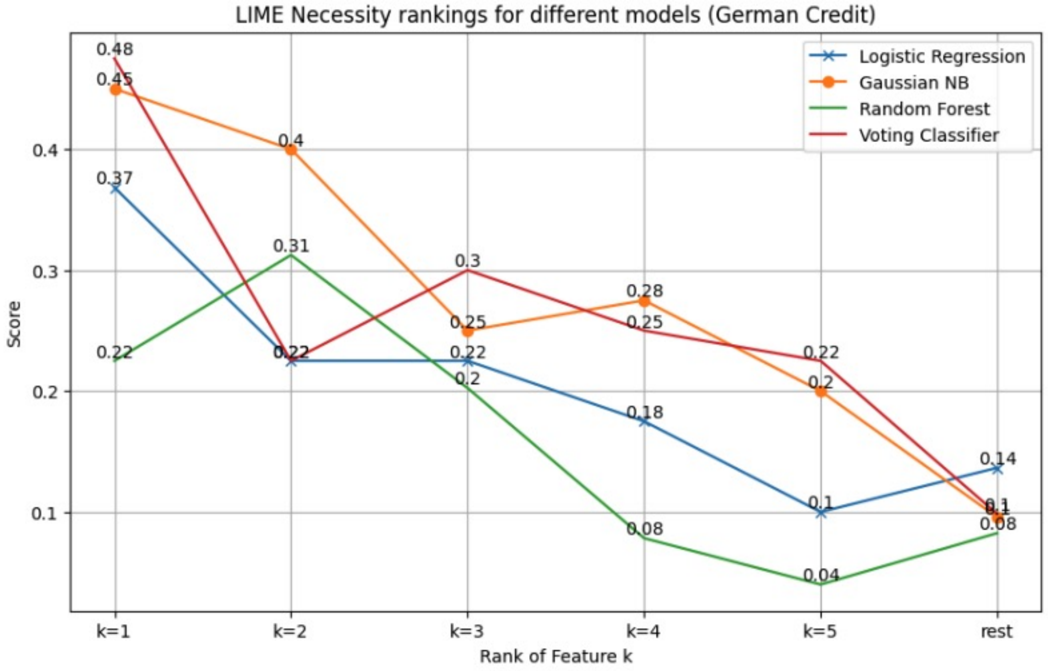


(a) LIME - Sufficiency

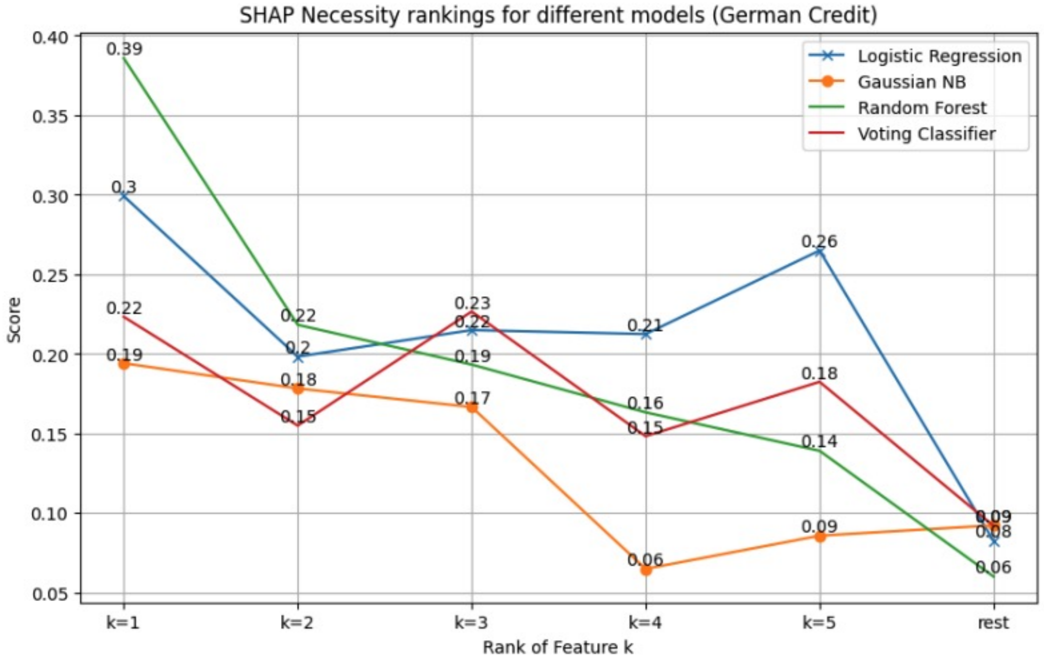


(b) SHAP - Sufficiency

Analysis of LIME and SHAP on German Credit Dataset



(a) LIME - Necessity



(b) SHAP - Necessity