Accuracy-Explainability tradeoff by explainable AI for complex ML model

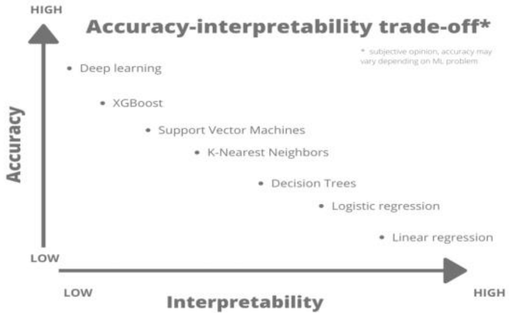
Poushali Sengupta (M.Sc),
Prof. Yan Zhang,
Prof. Sabita Maharjan, Prof. Frank Eliassen.

Department of Informatics ,
University of Oslo

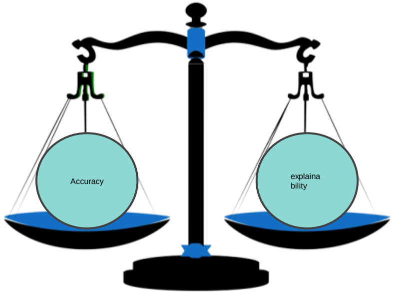
Date 02.11.2022

This presentation will cover the concepts of XAI, Existing Problems and our aim

What is Explainable AI (XAI)?



Existing Problem

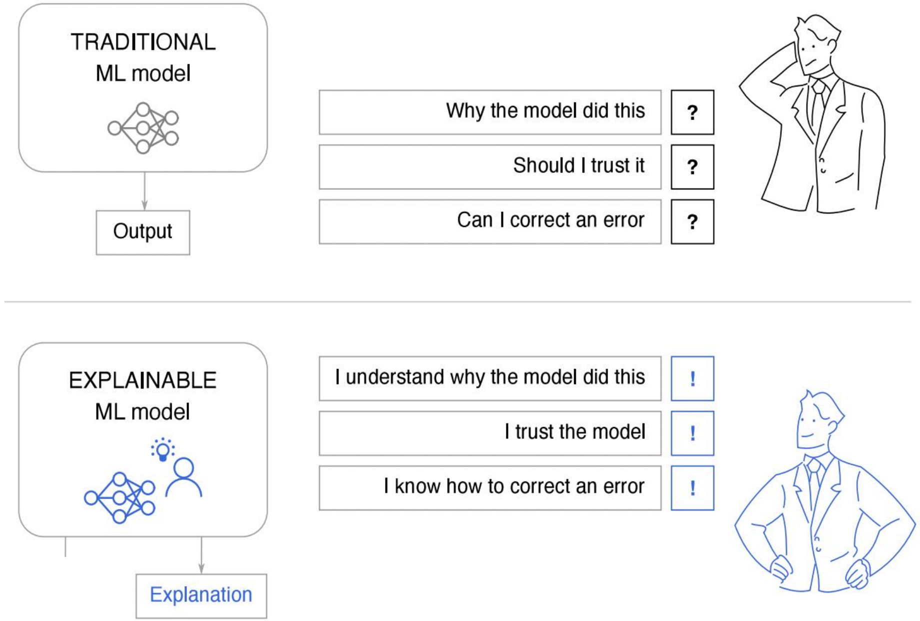


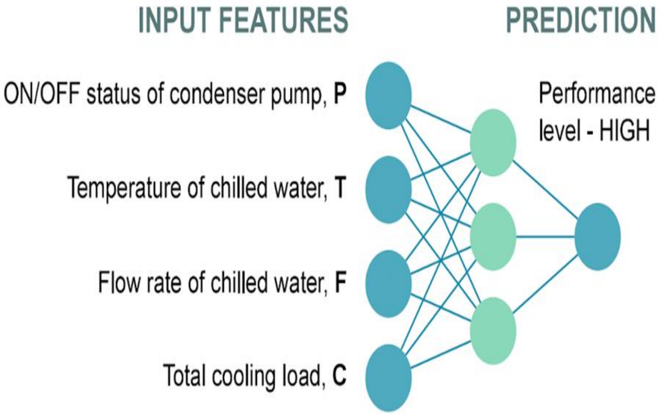
Our Aim

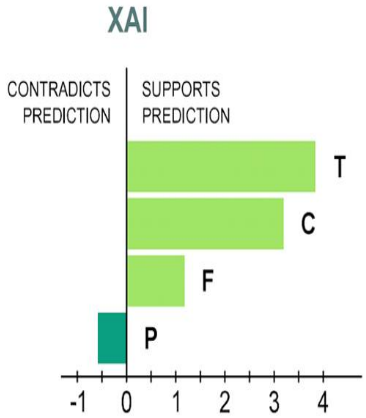
What is Explainable AI (XAI)?

Interpret the reason behind the **"Black Box"** of the Machine Learning algorithm.

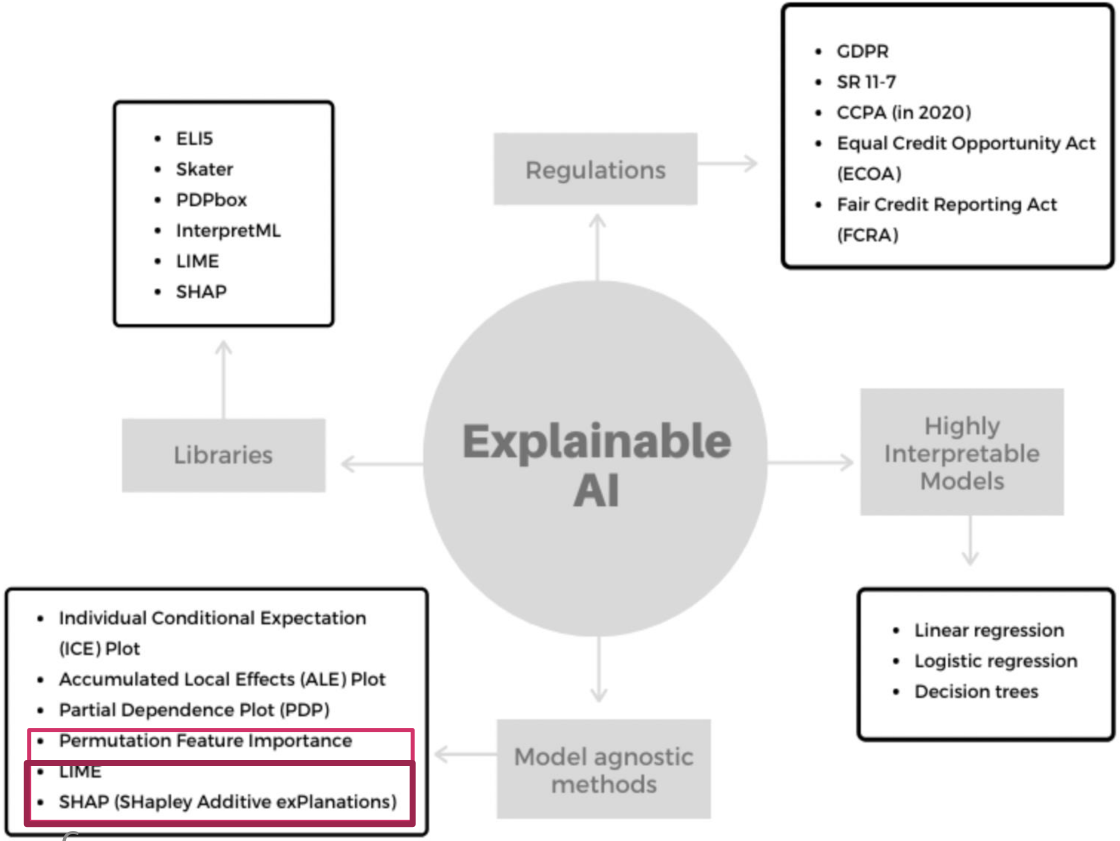
- Answers the question **"Why".**
- Explain the reason behind every **"decision".**
- Better understanding of the model to improve **"Accuracy"** and **"Trust".**







Existing Algorithms and Problems



Challenges

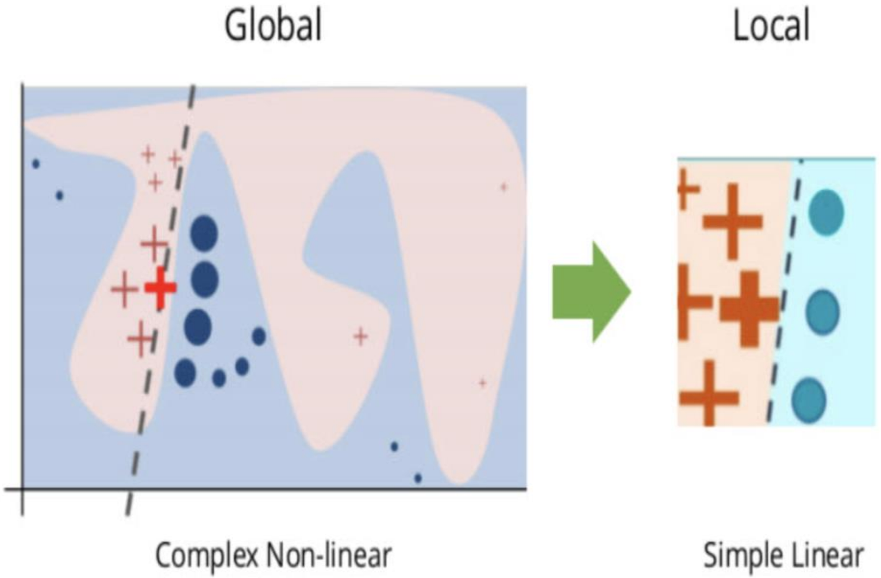
- Standardization
- Evaluation metrics
- Users
- Recommendations
- Trade Offs: Accuracy-Performance
- Privacy and Security

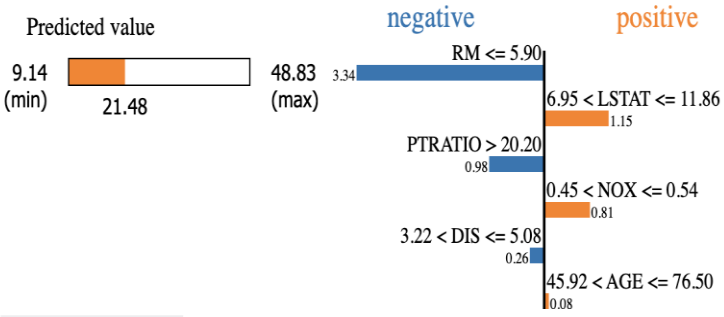
Focus

LIME: Local Interpretable Model-Agnostic Explanations

- **What is LIME?**
  - Explains individual predictions by building simple, interpretable models around each prediction.
- **How it Works:**
  - Perturbs the data and generates new samples.
  - Trains a simple model (e.g., linear regression) locally to explain the prediction.
- **Advantages:**
  - Works with any machine learning model (model-agnostic).
  - Provides easily understandable explanations for individual predictions.
- **Limitations:**
  - Can be unstable: Small changes in data can result in different explanations (uncertainty).
  - Struggles with complex models: May oversimplify the behavior of highly complex models, leading to less reliable explanations.

LIME: Local Interpretable Model-Agnostic Explanations

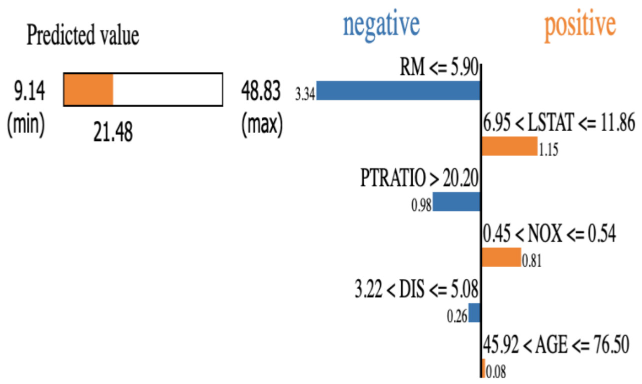




Feature	Value
RM	5.83
LSTAT	8.47
PTRATIO	21.00
NOX	0.54
DIS	4.50
AGE	56.50

LIME Prediction

- **Prediction:**
  - The model predicts a value of **21.48**, which falls between **9.14** (minimum) and **48.83** (maximum).
- **Feature Impact:**
  - LIME shows which features push the prediction up or down:
    - Blue bars (negative): Features that lower the prediction.
    - Orange bars (positive): Features that increase the prediction.
- **Examples:**
    - The number of rooms (RM) reduces the prediction by 3.34 (blue bar).
    - The level of LSTAT increases the prediction by 1.15 (orange bar).
- **Feature Table:**
  - Shows the actual values for key features that LIME uses to explain this prediction.



Feature	Value
RM	5.83
LSTAT	8.47
PTRATIO	21.00
NOX	0.54
DIS	4.50
AGE	56.50

What is SHAP? (SHapley Additive exPlanations)

- **Definition:**
  - SHAP helps us understand how much each feature in a model (like age, income, etc.) is contributing to a prediction (like the price of a house).

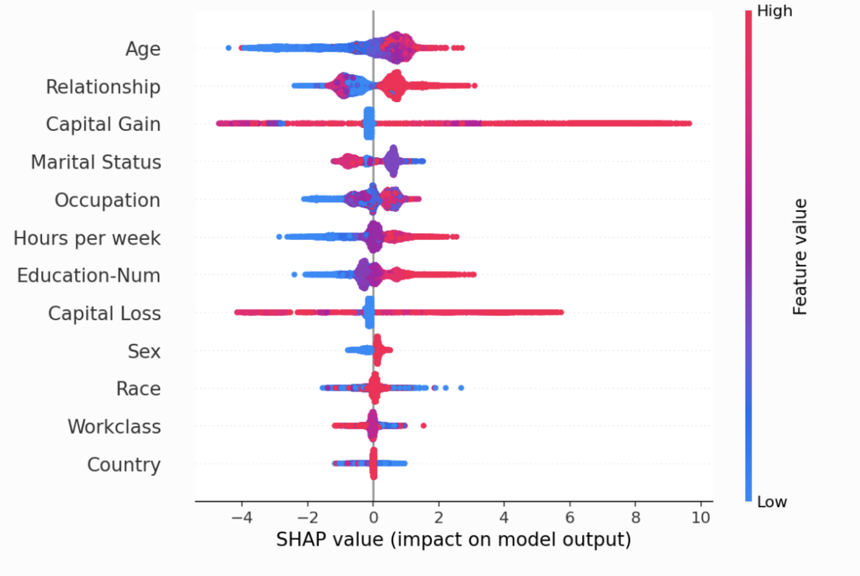
- **How It Works**
- **Shapley Values:**
  - SHAP looks at every feature one by one and checks how much it changes the prediction if we add or remove that feature.
- **Explains Each Prediction:**
  - It can explain both why the model made a specific prediction and how important each feature is for all predictions.

Problem With SHAP

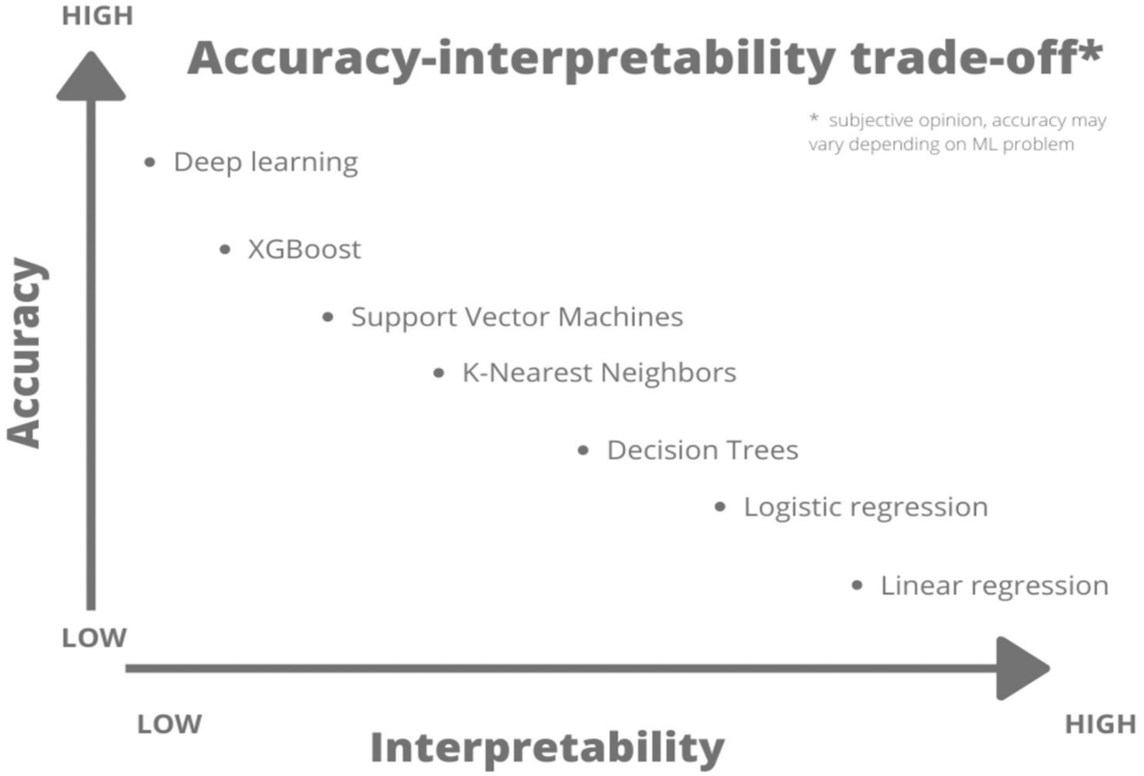
- **Feature Uncertainty:** SHAP doesn't handle correlated features well, leading to uncertain or inaccurate explanations when features depend on each other.
- **Accuracy vs. Explainability in Complex Models:** In very complex models, SHAP explanations can be too detailed and harder to understand, making it difficult to balance accuracy with clear explanations.
- **Needs a Lot of Computer Power**

SHAP Analysis For Income Prediction

- Age: Higher age increases income predictions.
- Relationship: Strong influence, both positive and negative.
- Capital Gain: High gain significantly raises income.
- Marital Status: Some statuses link to higher incomes.
- Education-Num: More education = higher income.
- Hours per Week: More hours worked, higher income.
- Other Features:Capital loss, sex, race, workclass, and country have smaller impacts.



More complex model hard to explain



Aim: Develop an alternative approach to balance Accuracy-Interpretability-Privacy for complex model

ExCIR (Explainability through Correlation Impact Ratio)

Accuracy Part

Light Weight Environment

- Secure the output accuracy
- Securing accuracy of explainability

Explainability part

Novel Explainability method

- **Introducing Correlation Impact Ratio for explainability**
- **Reduce computaional complexity**
- **Address feature uncertainity**
- **Work both with dependent and indepndent features**

How ExCIR Works:

$$\eta_{f_i} = \frac{n\left[\left(\hat{x_{f_i}} - \hat{x_{f_i y}}\right)^2 + \left(\hat{x_y} - \hat{x_{f_i y}}\right)^2\right]}{\sum_j \left(x_{f_i} j - \hat{x_{f_i y}}\right)^2 + \sum_j \left(x_{yj} - \hat{x_{f_i y}}\right)^2}$$

= Pertial CIR

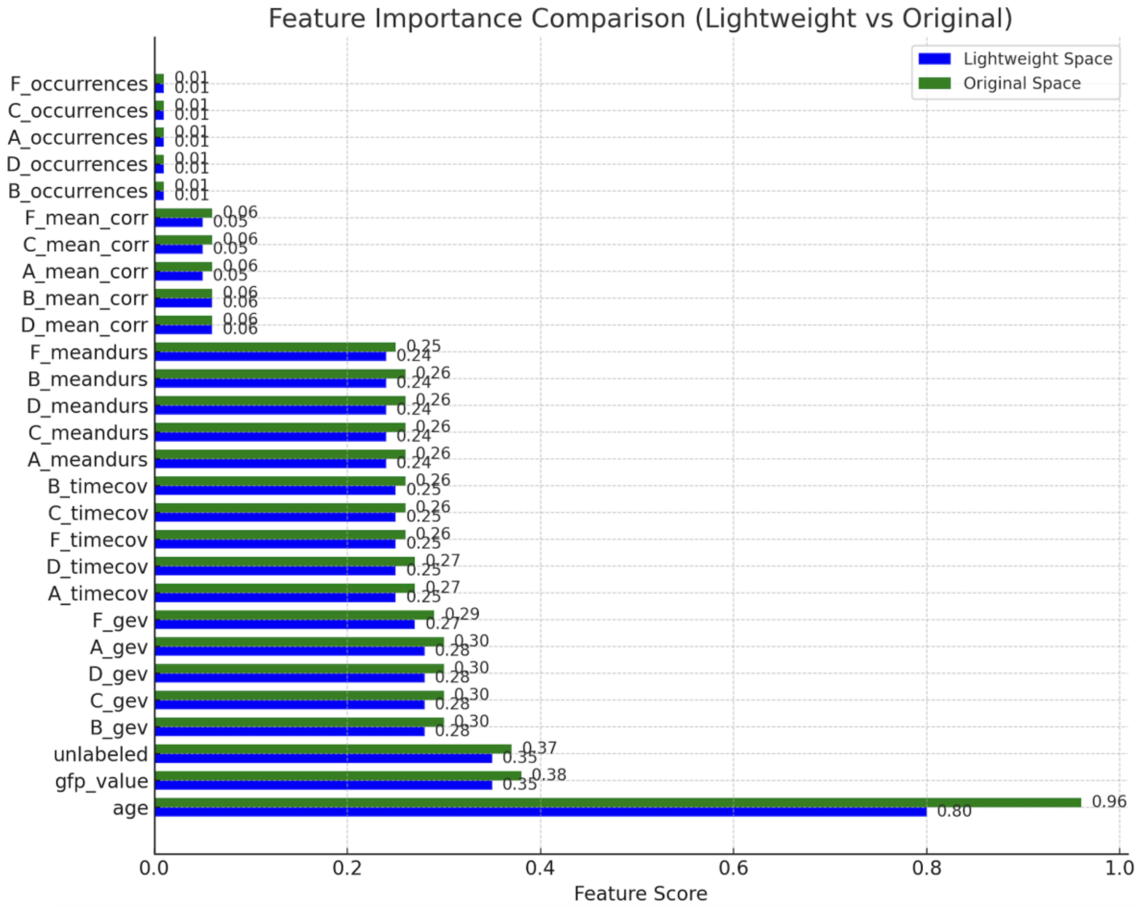
. Then, MCIR is: $C(\underset{\sim}{Y'}; \underset{\sim}{f_i} | \phi \subseteq \{||F||^{k\times n'} - \underset{\sim}{f_i}\}; i \neq j)) =$

$$\frac{I(\underset{\sim}{Y'}; \underset{\sim}{f_i} | \phi \subseteq \{||F||^{k\times n'} - \underset{\sim}{f_i}\})}{I(\underset{\sim}{Y'}; \underset{\sim}{f_i} | \phi \subseteq \{||F||^{k\times n'} - \underset{\sim}{f_i}\}) + I(\underset{\sim}{Y'}, \underset{\sim}{f_1}, \underset{\sim}{f_2}, \ldots \underset{\sim}{f_{i-1}}, \underset{\sim}{f_i}, \underset{\sim}{f_{i+1}}, .., \underset{\sim}{f_k})}$$

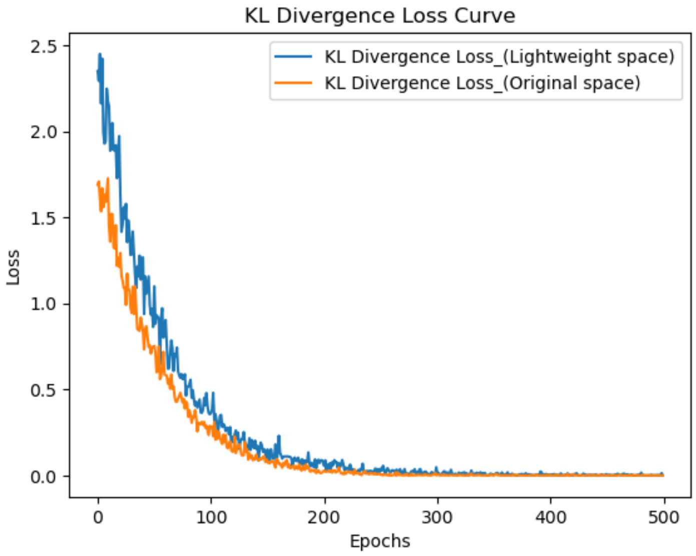
= Mutual CIR

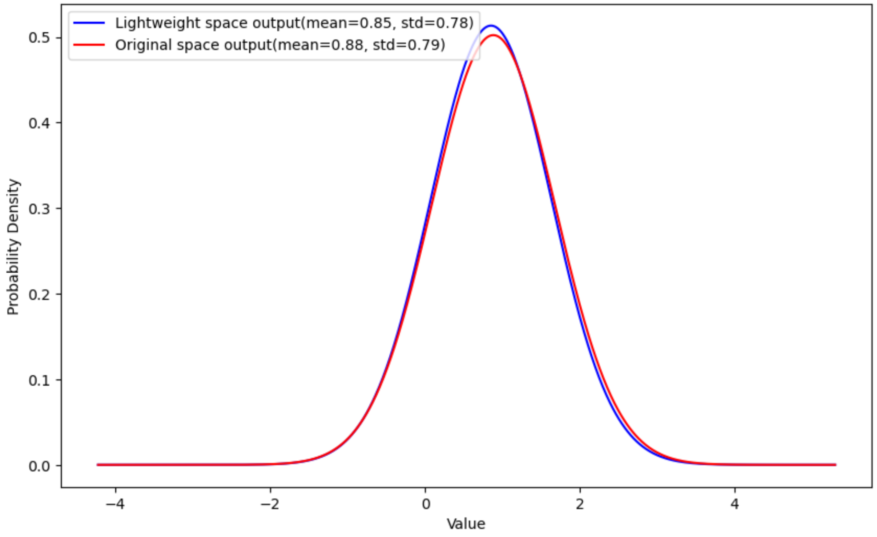
- Builds a **lightweight Environment** to simplify computation while retaining the structure of the original model.
- Use CIR score to explain feature impact
- It uses **Shannon entropy** to measure uncertainty in feature contributions.
- Ensures accuracy by aligning feature-output relationships between the original and lightweight models.

Explainability Consistancy



Accuracy of Orginial and Lightweight Model





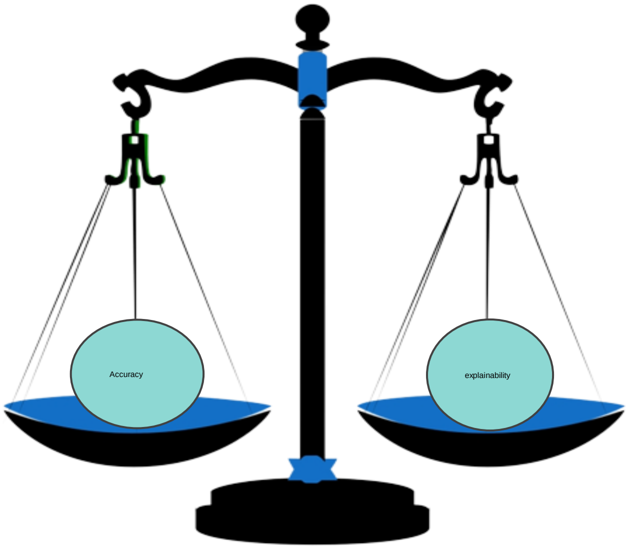
Results and Contribution

Main Contributions:

- Introduced a **novel metric** (CIR) to quantify feature importance, even with feature dependencies.
- Developed a framework that maintains the trade-off between explainability and accuracy.

Theoretical Results:

- Proven that ExCIR preserves model accuracy while improving interpretability.
- Ensures consistent feature ranking with minimal distortion, regardless of feature interdependence.



Question?

UiO : Universitetet i Oslo