Explainable AI

By Keshav, Khushi, Maissa, Oishani, Shivani

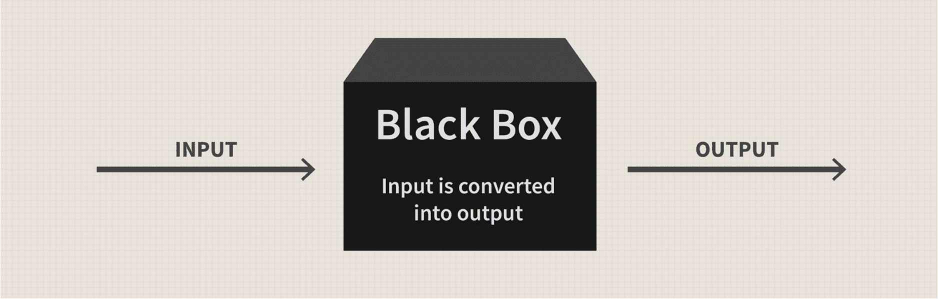
Introduction to Explainable AI

What is XAI?

XAI (Explainable Artificial Intelligence) is about making AI models more understandable to humans.

It helps make AI less of a black box as it helps explain how AI models make decisions so people can trust and understand them.

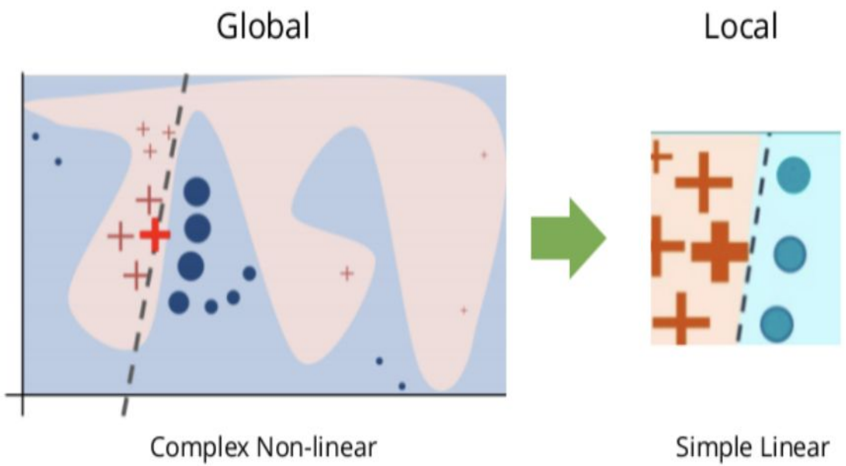
It also helps find mistakes and build trust among users.



Techniques

LIME

LIME (Local Interpretable Model-Agnostic Explanations)



How LIME works

Step 1: Pick a Data Point to Explain

Step 2: Create Small Changes (Perturbation)

Step 3: Observe Model Predictions

Step 4: Build a Simple (Interpretable) Model

Step 5: Show the Explanation

Example

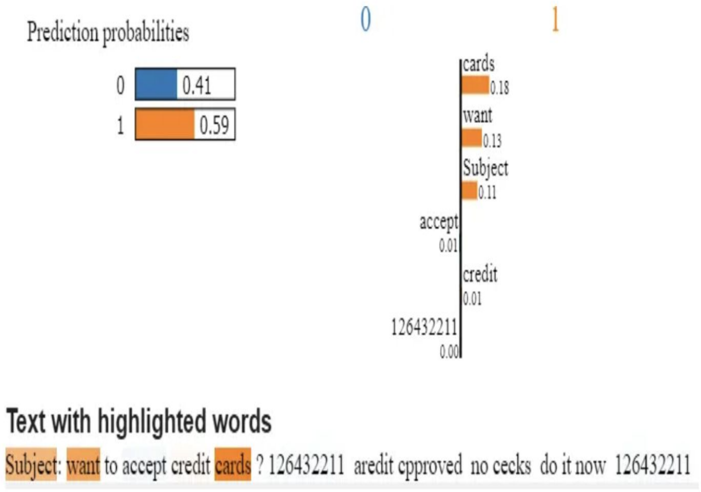
Email:

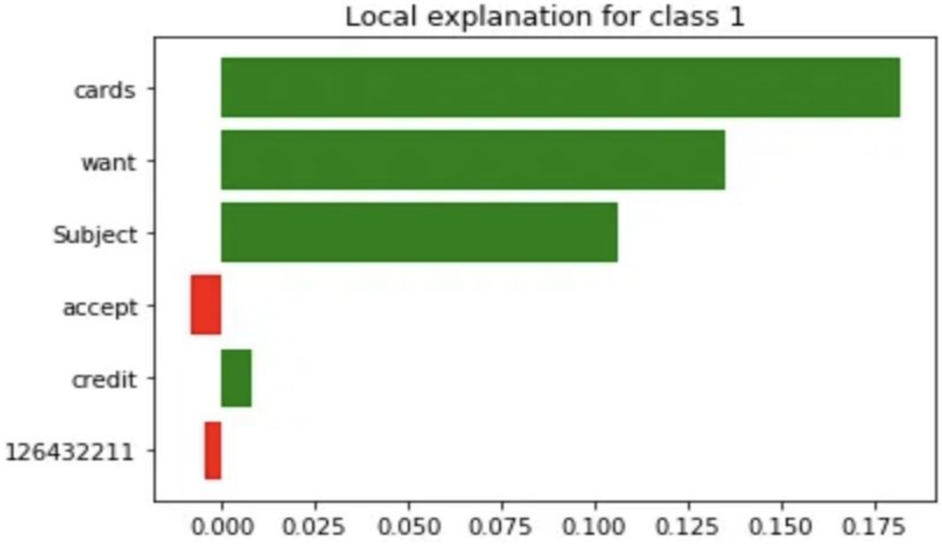
Subject: Want to accept credit cards? 126432211 aredit cpproved cecks do it now 126432211

Example walkthrough

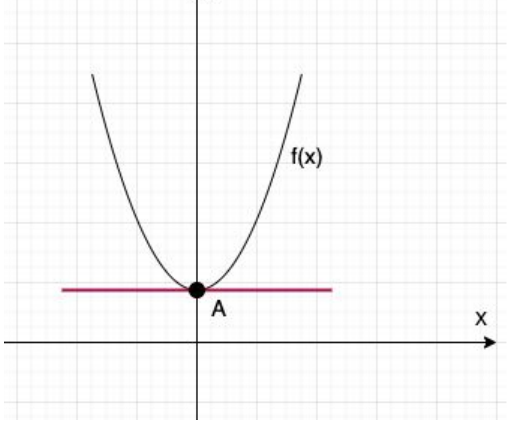
Concept:

- LIME provides a **visual breakdown** of which features (words) were most important for the prediction.





LIME: Mathematical Explanation



f is the black-box model.

g is the interpretable model.

π_x is a weighting function giving more importance to samples close to x.

L(f,g,π_x) is the loss function ensuring g mimics f locally.

Ω(g) ensures simplicity (e.g., limiting the number of features).

Mathematically:

$$\hat{g} = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

LIME: Mathematical Explanation

We will walk through **how LIME constructs the explanation model** using a simple **linear regression model**.

We want to explain why a black-box model $f(x)$ made a specific prediction for an instance x.

- **Goal:** Find a simple, interpretable model $g(z)$ that approximates $f(x)$ locally.
- **Assumption:** $g(z)$ is a **linear model** trained using perturbed samples.

LIME: Mathematical Explanation

Example: Predicting House Prices

Imagine we have a black-box model $f(x)$ that predicts house prices based on:

- Square Footage ($x1$)
- Number if bedrooms ($x2$)

We want to explain why the model predicted $300,000 for a house with:

$x1 = 2000$ sqft

$x2 = 3$ bedrooms

LIME: Mathematical Explanation

Step-1: Generate Perturbed Samples (z)

Sample	x_1 (sqft)	x_2 (bedrooms)
Original	2000	3
z_1	1900	3
z_2	2100	3
z_3	2000	4
z_4	2000	2

LIME: Mathematical Explanation

Step-2: Get predictions from the black-box model

Sample	x_1 (sqft)	x_2 (bedrooms)	$f(z)$ (Price Prediction)
Original	2000	3	**300,000**
z_1	1900	3	290,000
z_2	2100	3	310,000
z_3	2000	4	315,000
z_4	2000	2	285,000

LIME: Mathematical Explanation

Step-3 Compute the weights

We assign weights to each sample using the kernel function: $\pi_x(z) = e^{-\frac{D(x,z)^2}{\sigma^2}}$

Where D(x,z) is the euclidean distance between x and z

Sample	$D(x, z)$	$\pi_x(z)$ (Weight)
z_1	100	0.9
z_2	100	0.9
z_3	1	1.0
z_4	1	1.0

LIME: Mathematical Explanation

Step-4 Lime fits the local linear model

$$g(x) = w_0 + w_1 x_1 + w_2 x_2$$

We express this as a matrix equation: Xw = y

Sample	x_1 (sqft)	x_2 (bedrooms)	Predicted Price y
z_1	1900	3	290000
z_2	2100	3	310000
z_3	2000	4	315000
z_4	2000	2	285000

LIME: Mathematical Explanation

Step-4 Lime fits the local linear model

$$X = \begin{bmatrix} 1 & 1900 & 3 \\ 1 & 2100 & 3 \\ 1 & 2000 & 4 \\ 1 & 2000 & 2 \end{bmatrix}, \quad w = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix}, \quad y = \begin{bmatrix} 290000 \\ 310000 \\ 315000 \\ 285000 \end{bmatrix}$$

Sample	x_1 (sqft)	x_2 (bedrooms)	Predicted Price y
z_1	1900	3	290000
z_2	2100	3	310000
z_3	2000	4	315000
z_4	2000	2	285000

LIME: Mathematical Explanation

Step-4 Lime fits the local linear model

$$w = (X^T X)^{-1} X^T y$$

Compute $X^T X$

$$X^T X = \begin{bmatrix} 4 & 8000 & 12 \\ 8000 & 32000000 & 48000 \\ 12 & 48000 & 38 \end{bmatrix}$$

LIME: Mathematical Explanation

Step-4 Lime fits the local linear model

$$w = (X^T X)^{-1} X^T y$$

Compute $X^T y$

$$X^T y = \begin{bmatrix} 1200000 \\ 4800000000 \\ 7400000 \end{bmatrix}$$

LIME: Mathematical Explanation

Step-4 Lime fits the local linear model

$$w = (X^T X)^{-1} X^T y$$

Compute $(X^TX)^{-1}$

$$\begin{bmatrix} \frac{1}{2} & -\frac{1}{8000} & 0 \\ -\frac{1}{8000} & -\frac{1}{2720000000} & \frac{3}{68000} \\ 0 & \frac{3}{68000} & -\frac{1}{34} \end{bmatrix}$$

LIME: Mathematical Explanation

Step-4 Lime fits the local linear model

$$w = (X^T X)^{-1} X^T y$$

Solving for w by multiplying the inverse with $X^T y$

$$w = \begin{bmatrix} 55000 \\ 100 \\ 15000 \end{bmatrix}$$

LIME: Mathematical Explanation

Step-4 We fit the Local linear model

$$g(x) = w_0 + w_1 x_1 + w_2 x_2$$

Using least squares regression, we solve for w0, w1 and w2:

$$w_1 = 100$$

$$w_2 = 15,000$$

$$w_0 = 50,000$$

LIME: Mathematical Explanation

Step-5 Interpret the explanation

$$\text{Price} = 50,000 + 100 \times \text{Square Footage} + 15,000 \times \text{Bedrooms}$$

From g(x)

Each extra sqft increases price by $100 (positive effect).

Each extra bedroom increases price by $15,000 (stronger effect).

Neighborhood size (σ) impacts which samples contribute to the weights.

LIME: Mathematical Explanation

σ → **Neighborhood size**, controls how far samples are considered "local"

It determines how quickly weights drop as distance increases.

$$\pi_x(z) = e^{-\frac{D(x,z)^2}{\sigma^2}}$$

LIME: Mathematical Explanation

How does changing σ values change the weight?

$$\pi_x(z) = e^{-\frac{D(x,z)^2}{\sigma^2}}$$

$$\hat{g} = \arg\min_{g \in G} \sum_{i=1}^{m} \pi_x(z_i)(f(z_i) - g(z_i))^2 + \Omega(g)$$

LIME: Mathematical Explanation

Case 1: Small σ=50 (Highly Local)

$$\pi_x(z) = e^{-\frac{D(x,z)^2}{50^2}}$$

Sample	Distance $D(x, z)$	$\pi_x(z)$
z_1	100	$e^{-4} \approx 0.018$
z_2	100	$e^{-4} \approx 0.018$
z_3	1	$e^{-0.0004} \approx 0.9996$
z_4	1	$e^{-0.0004} \approx 0.9996$

Effect: Only the closest points (z3, z4) have high weights. Distant points (z1,z2) are nearly ignored

LIME: Mathematical Explanation

Case 2: Medium σ=100 (Balanced Neighborhood)

$$\pi_x(z) = e^{-\frac{D(x,z)^2}{100^2}}$$

Sample	Distance $D(x, z)$	$\pi_x(z)$
z_1	100	$e^{-1} \approx 0.367$
z_2	100	$e^{-1} \approx 0.367$
z_3	1	$e^{-0.0001} \approx 0.9999$
z_4	1	$e^{-0.0001} \approx 0.9999$

Effect: Now, distant points (z1,z2) still contribute, but with **less weight** than closer points.

LIME: Mathematical Explanation

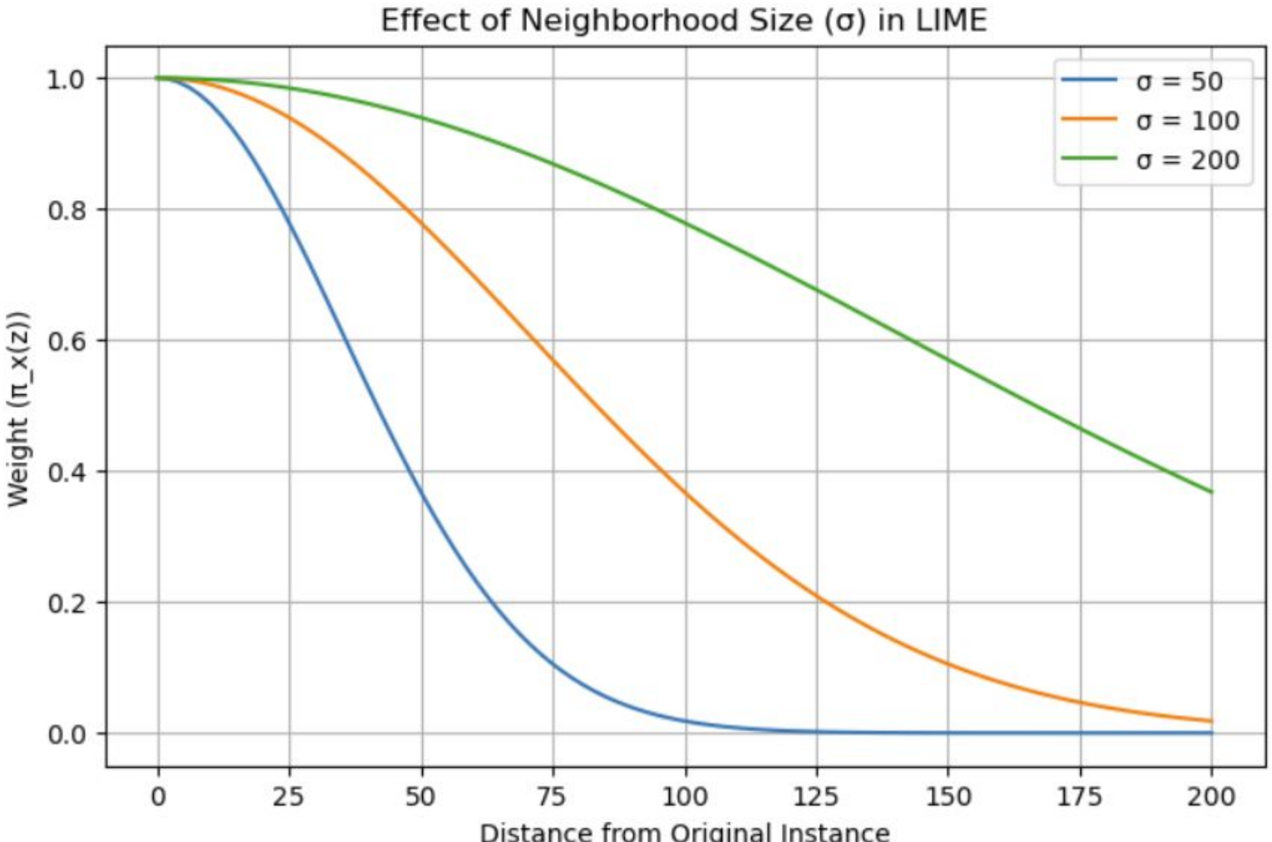
Case 3: Large σ=200 (Broad Neighborhood)

$$\pi_x(z) = e^{-\frac{D(x,z)^2}{200^2}}$$

Sample	Distance $D(x, z)$	$\pi_x(z)$
z_1	100	$e^{-0.25} \approx 0.779$
z_2	100	$e^{-0.25} \approx 0.779$
z_3	1	$e^{-0.000025} \approx 0.9999$
z_4	1	$e^{-0.000025} \approx 0.9999$

Even distant points (z1,z2) get high weights, meaning LIME looks at a **larger neighborhood**.

LIME: Mathematical Explanation



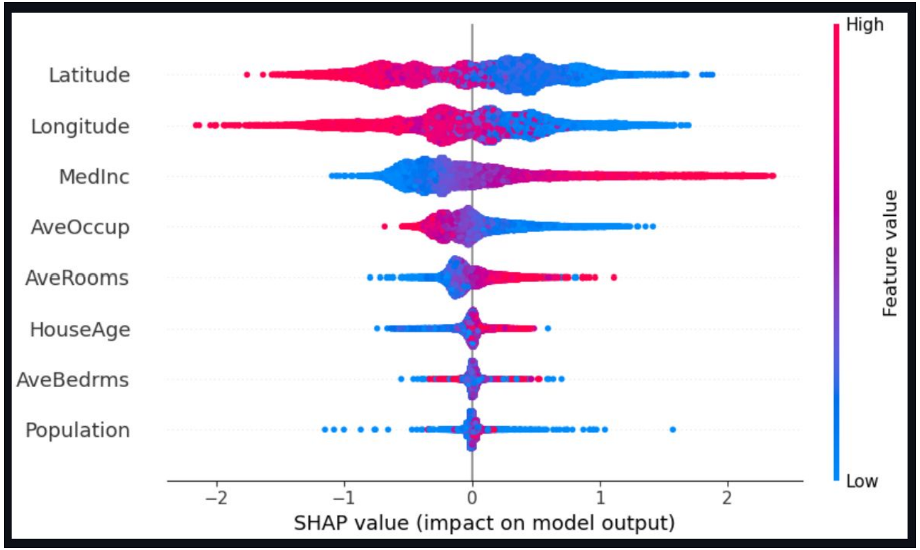
Example Python Implementation of LIME

LIME

- https://github.com/marcotcr/lime
- https://github.com/marcotcr/lime/blob/master/doc/notebooks/Tutorial%20-%20Image%20Classification%20Keras.ipynb (Demonstration in Class)

SHAP (SHapley Additive Explanations)

Explains the impact of each feature on a machine learning model’s predictions using principles from game theory.



How SHAP Works

Step 1: Baseline - Start with the average model prediction

Step 2: See how each feature changes the prediction from the baseline

Step 3: Coalitions - Test different combinations of features

Step 4: Assign SHapley Values to each feature based on its impact

Step 5: Visualize feature contributions with plots

Example: 2-Player Game by Connor O’ Sullivan

Game Details

- Goal: Prize money
  - Each Player is a **feature**
  - The team (2 players) is a **coalition**
  - **Marginal Contribution** is each player’s contribution
  - **SHAP Value** is the weighted average of each player’s contribution

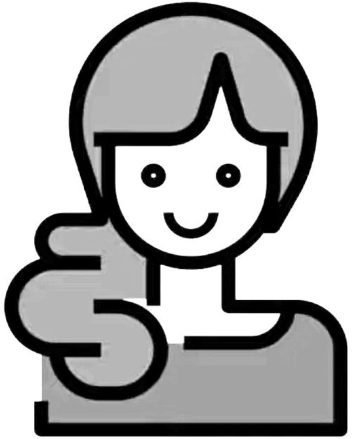
1st Place: $10,000

2nd Place: $7,500

3rd Place: $5,000

Team Prize: $10,000

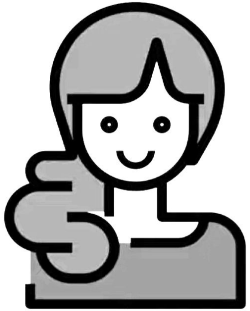
Player 1



Player 2

Coalitions

$$C_{12} = 10{,}000$$
$$C_{1} = 7{,}500$$
$$C_{2} = 5{,}000$$
$$C_{0} = 0$$



$$C_{12} - C_{2} = 5{,}000$$
$$C_{1} - C_{0} = 7{,}500$$

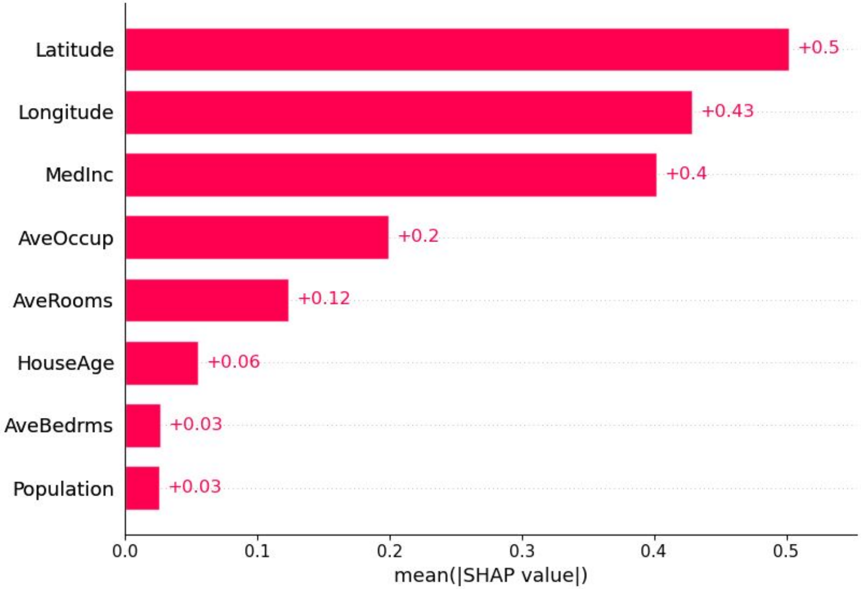
$$(5{,}000+7{,}500)/2$$
$$= \$6{,}250$$

$$C_{12} - C_{1} = 2{,}500$$
$$C_{2} - C_{0} = 5{,}000$$

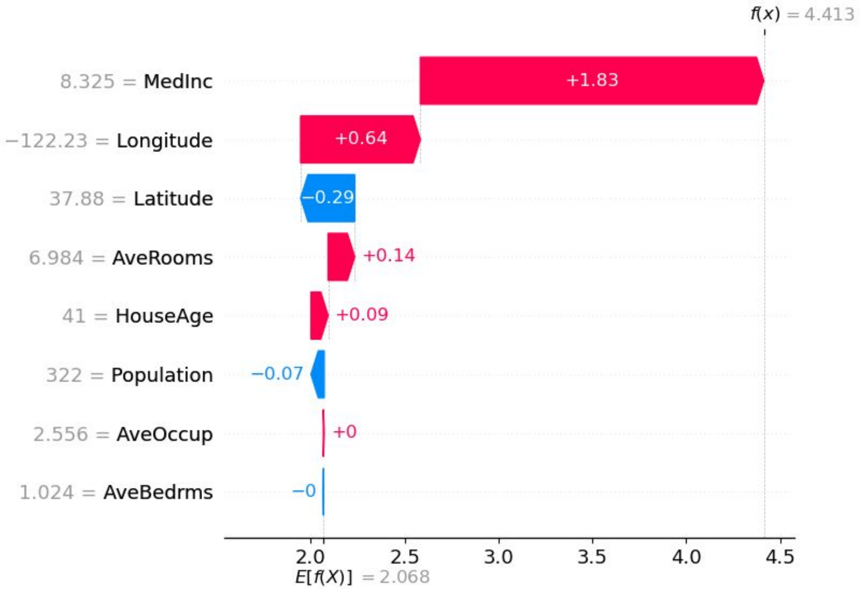
$$(2{,}500+5{,}000)/2$$
$$= \$3{,}750$$

Types of Visualizations

❖ Bar Plot



❖ Waterfall Plot



SHAP: Mathematical Explanation

N is the set of all features.

S is a subset of features excluding i.

f(S) is the model's prediction using only features in S.

The fraction is a weighting factor ensuring fair distribution.

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \underbrace{\frac{|S|!(|N| - |S| - 1)!}{|N|!}}_{\text{weight}} \underbrace{[f(S \cup \{i\}) - f(S)]}_{\text{marginal contribution of feature i}}$$

SHAP vs LIME

SHAP

- SHAP is based on game theory and provides more accurate, globally consistent attributions.
- SHAP is computationally expensive because it considers all possible feature combinations.
- SHAP can explain individual predictions and provide a global view of feature importance.

LIME

- LIME provides a simple, interpretable model (like linear regression) for local explanations.
- LIME is faster since it approximates the model locally.
- LIME focuses on a single instance at a time with a locally trained model

When to use SHAP vs LIME?

SHAP

- Detailed, consistent explanations are required
  - SHAP offers mathematically consistent and stable feature attributions
- Global & local insights are required
  - Useful when the overall behavior of the model is important
- Computational cost is not a limitation
  - Since all possible feature coalitions are computed, SHAP is expensive

LIME

- Quick, approximate explanations are required
  - LIME offers fast approximations by perturbing input data and fitting a simpler model locally
- Working with high-dimensional data
  - Since not all feature coalitions are computed, LIME is more scalable
- You simply prefer a model-agnostic method
  - LIME is independent of the underlying model
  - Useful when comparing multiple black-box models where interpretability is crucial
- Linear Model biases - good for quickness, bad for subtle interactions!
  - assumes feature independence, unlike SHAP

$$g(x) = w_0 + w_1 x_1 + w_2 x_2$$

Applications and Ethics

Exercise: Challenges and Benefits of XAI in Industry

Applications and Ethics

What are possible challenges of using XAI?

Applications and Ethics

What are possible challenges of using XAI?

Think about:

1. Complex models
2. Hacking
3. Priorities of companies in industry
4. Various kinds of industry applications and data privacy

Applications and Ethics

Challenges

1. Accuracy
   a. best with deep natural networks
   b. complicated for basic XAI
2. Hackability
   a. models with XAI more easily reverse engineered
   b. important parameters identified
3. Data Privacy
   a. privacy concerns for financial and health data
   b. data used for large models might be proprietary
4. Intellectual Property
   a. company algorithms designed to be proprietary & profitable
   b. latest AI regulations keep changing

Applications and Ethics

What are the benefits of using XAI?

Applications and Ethics

What are the benefits of using XAI?

Think about:

1. Biases
2. Improving models and error rates
3. Legality and interpretability
4. Users of AI

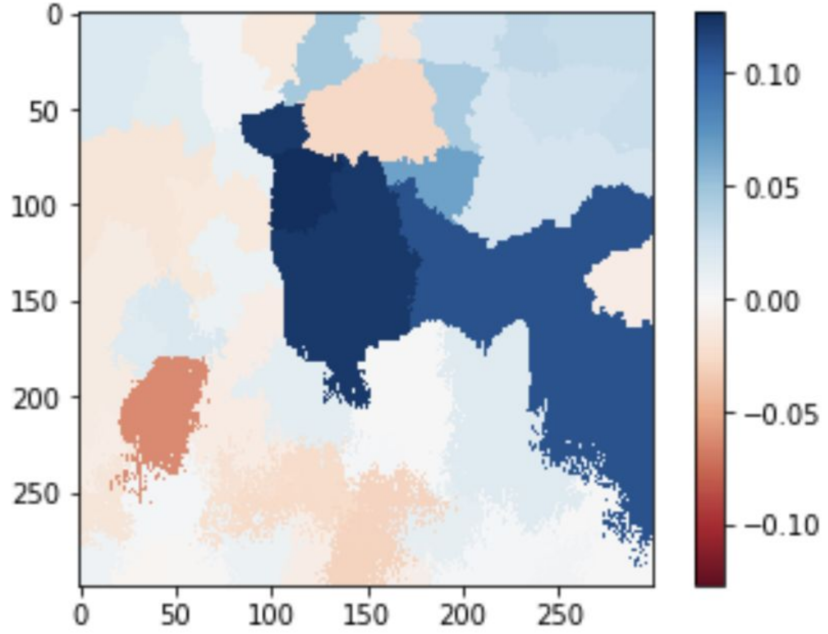
Applications and Ethics

Benefits

1. Debiasing
   a. bias easier to detect and rectify
   b. using XAI can debias results & improve overall accuracy
2. Model Improvement
   a. factors of model to fine tune and improve
   b. errors inspected closely
3. User Trust
   a. algorithms become more interpretable and explainable
   b. users understand and trust them more
4. Risk and Compliance
   a. model explanations help legal compliance with AI regulations
   b. liability risks reduced by using XAI

Questions?

Thank You!



https://github.com/marcotcr/lime/blob/master/doc/notebooks/Tutorial%20-%20Image%20Classification%20Keras.ipynb