Demystifying Black-Box Models: A Study of Local Interpretable Model-Agnostic Explanations

Student Name: M B Thejesshwar
Roll Number: 24AI10035

Abstract

As Machine Learning models particularly deep neural networks and ensemble methods become increasingly complex they often sacrifice interpretability for accuracy. This black box nature poses significant challenges in critical domains such as healthcare and finance where understanding the decision making process is as vital as the prediction itself. This project explores **Local Interpretable Model Agnostic Explanations,** a technique that explains the predictions of any classifier in an interpretable and faithful manner by approximating it locally with an interpretable model. We demonstrate the efficacy of LIME on a synthetic non linear dataset, showing how it successfully isolates local decision boundaries where global linear models fail. The study confirms that LIME provides faithful local explanations to complex non linear boundaries. We further validate the approach by quantifying the stability of the explanations across multiple stochastic runs

Introduction and Motivation

The trade off between model accuracy and model interpretability is a fundamental tension in Machine Learning. Simple models like Linear Regression and Decision Trees are highly interpretable; we can easily understand how features contribute to the output. However these models often lack the capacity to model complex non linear relationships found in real world data.

Conversely complex models like Random Forests, Gradient Boosting Machines and Deep Neural Networks achieve state of the art performance but act as black boxes. For instance a Random Forest might aggregate the votes of 100 decision trees making it impossible for a human to trace the logic behind a specific classification.

The Need for Why

In high-stakes applications because the algorithm said so is insufficient.

- **Medical Diagnosis:** A doctor cannot trust a model diagnosing cancer without knowing which symptoms drove that decision.
- **Algorithmic Fairness:** We must ensure a model is not biased against protected attributes.

Literature Review: LIME vs. SHAP

Two primary frameworks dominate the field of Explainable AI LIME and SHAP.

- **SHAP** is based on cooperative game theory and provides a globally consistent explanation by calculating the marginal contribution of each feature. However it is often computationally expensive for large datasets.
- **LIME** focuses purely on **local fidelity**. It does not attempt to explain the model's behavior on the entire dataset but rather fits a simple model around a single prediction.

This project focuses on **LIME** due to its computational efficiency and its intuitive geometric interpretation.

Theoretical Framework

The core intuition behind LIME is that even if a decision boundary is highly non linear globally it looks linear if we zoom in close enough to a single point. This is analogous to calculus where a curve can be approximated by a tangent line at a specific point.

Mathematical Formulation

LIME generates an explanation by solving the following optimization problem:

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \quad \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

Where:

- **x**: The specific instance being explained
- **f**: The black-box model whose probability output we want to explain
- **g**: An interpretable model from a class G
- **pi_x**: A proximity measure that defines the locality around x
- **Omega(g)**: A complexity penalty For linear models this penalizes the number of non zero weights

The Loss Function

To find the best local approximation, LIME minimizes the weighted squared loss:

$$\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in \mathcal{Z}} \pi_x(z) \left(f(z) - g(z') \right)^2$$

Here, we sample perturbed instances z around the original point. The proximity kernel pi_x(z) weights these samples using an exponential kernel based on Euclidean distance D:

pi_x(z) = $exp(-D(x, z)^2/\sigma^2)$

This ensures that samples close to our original point x have a massive influence on the explanation g while points far away are ignored.

Feature Selection

To ensure the explanation is understandable to humans LIME typically uses **Lasso Regression** or **K Lasso** to select only the top K most important features effectively setting the weights of irrelevant features to zero.

Experimental Design

To rigorously demonstrate the advantage of LIME we designed an experiment using synthetic data where the ground truth geometry is known.

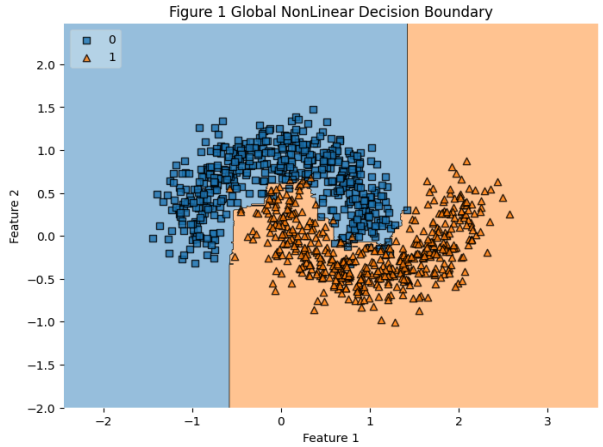
- **Dataset:** Two Moons
  - *Characteristics:* Non-linear, interleaving classes
  - *Sample Size:* 1000 samples
  - *Noise:* 0.2
- **Model:** Random Forest Classifier
  - *Reason for choice:* It is a classic black box model that easily handles non linearities
- **Comparison:** We visualize the global decision boundary versus the local explanation provided by LIME

Results and Analysis

LIME.ipynb

Global Decision Boundary

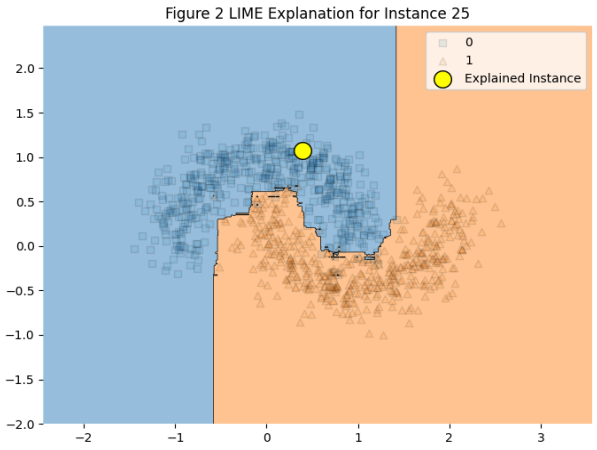
Figure 1 displays the decision surface learned by the Random Forest. As expected the model learns a highly non linear S shaped boundary to separate Class 0 from Class 1. A traditional global linear model would fail here achieving only 50-60% accuracy.

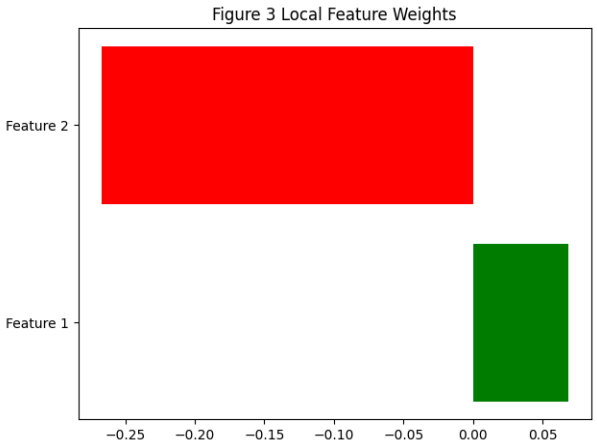


Local Explanation Analysis

We selected a specific test instance located at the edge of the upper moon. The Random Forest predicts this point belongs to Class 0 with high confidence.

Figure 2 shows the LIME explanation for this point. The visualization reveals the local linear boundary that LIME fits to the Random Forest's predictions.





Analysis of Weights: The feature importance weights returned by LIME for this instance are:

- **Feature 2:** Weight `-0.2669923174497469`
- **Feature 1:** Weight `0.06825130124871204`

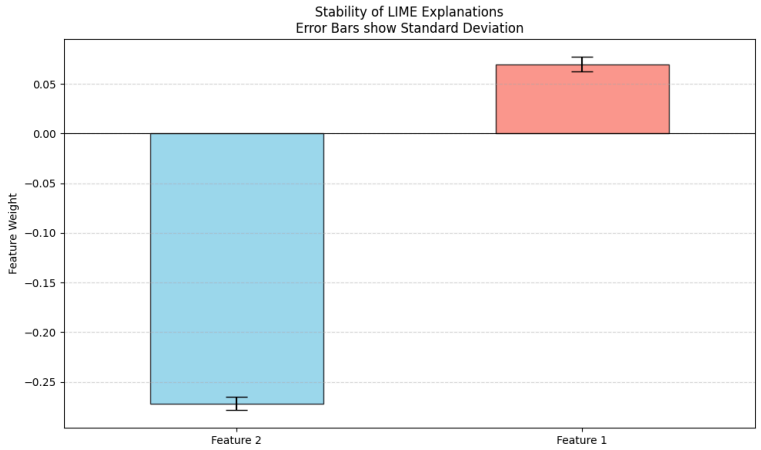
Interpretation: This specific weighting is physically meaningful. For this specific point on the moon moving *down* pushes the point deeper into the Red class while moving *up* pushes it toward the Blue class. LIME has correctly identified that **vertical position** is the dominant factor for classification *in this specific neighborhood* even though both features matter globally. This confirms that LIME successfully captured the local tangent of the non linear manifold.

Robustness and Hyperparameter Analysis

To validate the reliability of the LIME explanations we conducted two additional experiments focusing on stability and hyperparameter sensitivity.

Stability Analysis

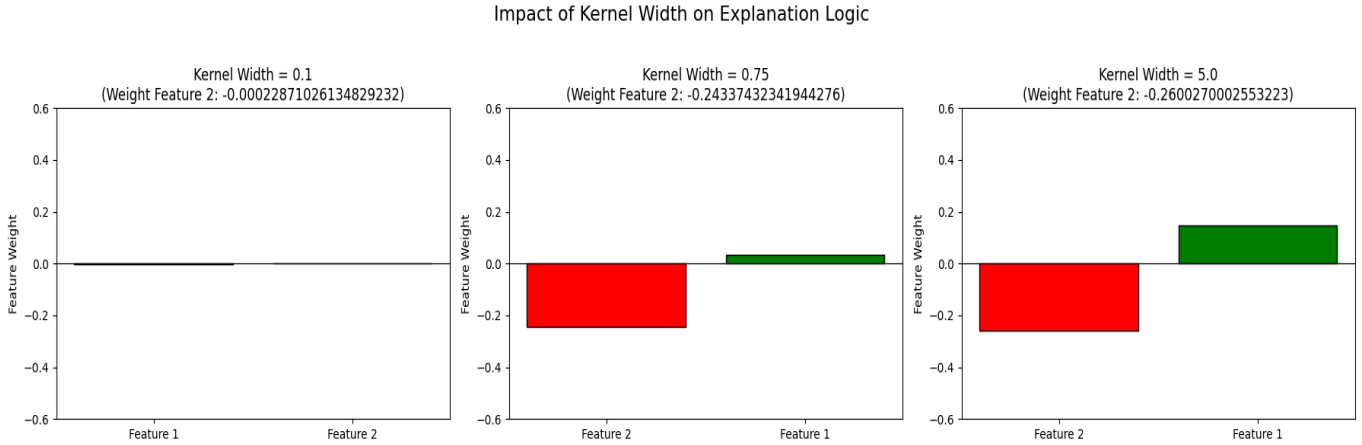
A common critique of LIME is its stochastic nature since it generates explanations by random sampling; different runs can theoretically yield different feature weights for the same instance. To quantify this risk we executed the LIME explainer **10 times** on the same test instance.



Observation: As seen in Figure the standard deviation of the feature weights is negligible. The explanation consistently identifies Feature 2 as the dominant negative factor confirming that the model's interpretation is stable and reproducible for this dataset.

Kernel Width Sensitivity

The locality of LIME is defined by the **Kernel Width**. We analyzed how varying sigma alters the explanation logic.



Observation:

- **sigma = 0.1:** The explanation becomes unstable overfitting to the nearest single neighbor.

- **sigma = 0.75:** Captures the local linear tangent effectively.
- **sigma = 5.0:** The explanation over smooths the decision boundary effectively reverting to a global average that fails to capture the non linear nuance of the Two Moons dataset.

Conclusion

This project demonstrated that complex black box models can be trusted if we interpret them locally. By applying LIME to a synthetic non linear dataset we showed that the algorithm effectively ignores the global complexity of the Random Forest and focuses on the immediate decision boundary near the instance of interest. The result is a simple linear explanation that faithfully represents the model's local behavior bridging the gap between high accuracy AI and human interpretability. Our stability analysis confirmed that the local explanations are robust with low variance across repeated trials provided an appropriate kernel width is selected.

References

"Why Should I Trust You?" | Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining
A unified approach to interpreting model predictions | Proceedings of the 31st International Conference on Neural Information Processing Systems
Interpretable Machine Learning