**Assignment: Feature attribution methods and their evaluation**

**Due: Nov 21, 2024**

***Note: I had permission from the professor for an extension of a few days after the date, so I’m submitting this assignment on the 23rd.***

**Objective**The goal of this assignment is to explore how various feature attribution methods provide explanations for model predictions and to assess their effectiveness using quantitative metrics.

**Dataset**

**Tabular Dataset:** Used the [Bank Marketing Dataset](https://www.openml.org/search?type=data&id=44234&sort=runs&status=active) from UCI on [OpenML](https://www.openml.org/). The dataset is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

**Model**

To make predictions on this dataset, a specific 2 layer simple feedforward neural network was implemented.

**Tasks**

1. **Model Training:** Prepared and trained the model on the chosen dataset.

A white background with black text

Description automatically generated

A screenshot of a computer

Description automatically generated

1. **Explainability Analysis:**
   1. Evaluated explanations generated by the following methods: LIME, Gradient, InputXGradient, and IntegratedGradients.
   2. Implemented several XAI models to monitor explainability:
      1. LIME – from [Github](https://github.com/marcotcr/lime). I did not find the one from Captum to work well, hence decided on using the original implementation from Github.
      2. GradientSHAP – from [Captum](https://captum.ai/api/gradient_shap.html)
      3. InputXGradients – from [Captum](https://captum.ai/api/input_x_gradient.html)
      4. IntegratedGradients – from [Captum](https://captum.ai/docs/extension/integrated_gradients)
   3. Confirmed the availability of these methods with the following command: **quantus.AVAILABLE\_XAI\_METHODS\_CAPTUM.**

Available XAI Methods: ['GradientShap', 'IntegratedGradients', 'DeepLift', 'DeepLiftShap', 'InputXGradient', 'Saliency', 'FeatureAblation', 'Deconvolution', 'FeaturePermutation', 'Lime', 'KernelShap', 'LRP', 'Gradient', 'Occlusion', 'LayerGradCam', 'GuidedGradCam', 'LayerConductance', 'LayerActivation', 'InternalInfluence', 'LayerGradientXActivation', 'Control Var. Sobel Filter', 'Control Var. Constant', 'Control Var. Random Uniform']

* 1. **Qualitative Analysis**

Generated Feature Importance Graphs for each explainable model.

**LIME:**

**With only 7 features.**

A close-up of a person's face

Description automatically generated

**With all 16 features.**

A screen shot of a computer

Description automatically generated

**A graph with a green and red bar graph

Description automatically generated**

**Integrated Gradients:**

A graph with blue squares

Description automatically generated

**InputXGradients:**

A graph with blue squares

Description automatically generated

**GradientSHAP:**

**A graph with blue squares

Description automatically generated**

**Observations**

1. **LIME**:
   * LIME focuses on individual, datapoints rather than larger regions of interest.
   * In a bank marketing dataset, this might be useful to understand the predictions made for a single person, but in a different domain such as images, it might be irrelevant.
   * In my case, LIME for a singular instance gave a strong probability score for it to predict ‘No’ for a bank term deposit based on just 5 random features.
   * Later I used 16 features to understand the impact on a different output and it still gave me a good explanation for which features are the most useful.
2. **IntegratedGradients**:
   * This method highlights broader, continuous regions, capturing global dependencies and relationships among various features in a tabular dataset
   * For instance, in this specific case, IG demonstrated **‘age’, ‘education’ and ‘month’ and ‘marital status’** as positive indicators to predict whether the product (bank deposit) would be subscribed or not.
   * Indeed, as a human insight, education and age holds significant importance in predicting whether an individual will make a bank term deposit. This is shown in the graph as positive values for **education** at **0.145** and **age** at **0.091**. The **most important feature** was the **month** feature with a attribution score of **0.467**
   * It effectively balances relevance across the entire object, making it suitable for comprehending the general significance of all features in the tabular dataset.
3. **InputXGradient**:
   * In an image, InputXGradient emphasizes **edges and fine details**. In the context of a **tabular dataset**, the method evaluates how changes in individual input features impact the model's output by computing the gradient of the output with respect to each input feature.
   * In my case, I have aggregated the impact into a single feature importance graph.
   * **Scale Dependency:** Inputx gradients can be influenced by the scale of input features, so normalization is often required. This is the reason for normalizing my dataset before applying XAI models on the dataset.
   * In my case, the graph is almost so **identical with IntegratedGradients**. However, the attribution values are slightly larger in comparison for some features.
   * Moreover, it solidifies education as a important predictor and confirms the findings of IntegratedGradients as both gave ‘education’, ‘age’ and ‘month’ higher significance.
   * All these features are important predictors in the real world of finance, thus the model is working well in predicting the output.
4. **Gradient (GradientSHAP)**:
   * Gradients focus on **edge cases** and highlight similar regions as InputXGradient but with **higher noise** and less smoothness.
   * Rightly so, in my case, GradientSHAP gave a **varying output** compared to the other three XAI models.
   * In this particular case, GradientSHAP used a normal baseline and the same parameters as the other model and yet, gave only **‘age’** a positive attribution score.

XAI methods such as LIME, IntegratedGradients, InputXGradient, and GradientSHAP provide insights into model predictions. LIME and IntegratedGradients highlight individual features, while InputXGradient and GradientSHAP emphasize edges and fine details. These methods demonstrate the significance of education, age, and month in predicting bank term deposit subscriptions.

Integrated Gradients excels in highlighting global explainability that captures critical object features. InputXGradient focuses on finer details, offering precise explanations but introducing slight noise. LIME, however, produced accurate point-based locally explainable graphs for feature importance. Gradient-based SHAP explanations are noisier and less interpretable than Integrated Gradients and InputXGradient. In comparison, InputXGradients and IntegratedGradients performed significantly better in explaining the model compared to the other methods.

***Note: For Quantitative analysis, my code was written for tabular data as I wanted to keep my assignment unique. and I didn’t find any supplemnraty material for using these specific metrics for tabular. All of them were mostly used for images online. I did try to do it for tabular which can be found in my previous github commits. Please look at them and let me know if I was on right track or not. I did ask Raghu and Dip but still couldn’t figure it out for tabular. I’ll be happy to add them after I figure out how to do them for tabular.***

* 1. **Quantitative Analysis**
     1. For each explanation method, evaluated the following metrics:
        1. **Faithfulness Correlation:** Measures how well the explanation aligns with the model's predictions. Link: <https://quantus.readthedocs.io/en/latest/docs_api/quantus.metrics.faithfulness.faithfulness_correlation.html>
        2. **Relative Input Stability:** Evaluates the robustness of explanations under slight input perturbations. The lower, the better. Link:  
           <https://quantus.readthedocs.io/en/latest/docs_api/quantus.metrics.robustness.html>

* + - 1. **Sparsity:** Assesses the simplicity of explanations by checking how many features contribute significantly. Link:   
         <https://quantus.readthedocs.io/en/latest/docs_api/quantus.metrics.complexity.sparseness.html>
    1. After computing these metrics, analyze the results. Consider the following questions:
       1. Which explanation method scores highest for faithfulness, stability, and sparsity?
       2. Is any method consistently the best across all metrics, or do trade-offs exist?
       3. Does the quantitative evaluation align with the qualitative observations?
  1. **Points to remember:**
     1. Always perform analyses on correctly classified samples to ensure meaningful and reliable insights.
     2. When calculating metrics, average the results over 500/1000 correctly classified samples to obtain robust quantitative evaluations.

**Submission**

Submit a report with your findings, including heat maps or tables, and a detailed analysis of each explanation method's effectiveness based on both qualitative and quantitative metrics.