**Project Description**

This project focuses on addressing a critical limitation of the original Local Interpretable Model-agnostic Explanations (LIME), which generates explanations for machine learning models by creating random perturbations of the input data and observing how the model’s predictions change. While LIME provides a useful framework for model interpretability, its reliance on **random perturbations** often results in:

• **Irrelevant inputs** which do not align with real-world data distributions.

• **Lack of domain-specific context**, making the explanations less meaningful for specialized applications.

• **Reduced stability and fidelity**, as explanations may vary significantly with slight changes in data or model behavior.  
A diagram of a diagram

Description automatically generated

**Original LIME Process:**

1. LIME perturbs the input data by randomly tweaking feature values (e.g., masking or changing data points).

A close-up of a paper

Description automatically generated

2. It trains a local, simpler model (e.g., linear regression) on the perturbed data to approximate the global model’s predictions.

3. The importance of features is derived from the weights of this local model, which is used to explain the original model’s decision for a specific input.

**Proposed Solution:**

To overcome these limitations, we propose an enhanced version of LIME by rethinking its **perturbation mechanism** and incorporating **domain-specific context** through fine-tuned Large Language Models (LLMs).

Key innovations include:

1. **Domain-Adaptive Fine-Tuning**:

• Fine-tuning LLMs, such as BERT or DeBERTa, using masked language modeling (MLM) for the specific dataset or application domain.

• This ensures the LLM understands the nuances of the data and generates perturbations that align with realistic patterns.

2. **Context Aware Perturbation Function**:

• Instead of random perturbations, we generate **context-aware textual descriptions** by masking random columns of data and using the fine-tuned LLM to predict plausible replacements.

• These meaningful perturbations lead to more interpretable and accurate explanations.

3. **Metrics Definition**:

• **Fidelity**: Ensuring the local explanations align closely with the global model’s behavior.

• **Stability**: Measuring the consistency of explanations when similar inputs are perturbed.

**Intended Output**:

For a tabular dataset in a healthcare application:

• Original LIME might randomly change patient age from 30 to 100, which is unrealistic and irrelevant.

• Enhanced LIME masks the age column and uses the fine-tuned LLM to predict a plausible replacement (e.g., 32), grounded in the domain context.

This improved approach not only enhances the quality of explanations but also makes them more actionable and reliable, benefiting both technical users (e.g., data scientists debugging models) and non-technical stakeholders (e.g., healthcare professionals interpreting model results).

**Technical Problems**

**Approach to the Problem**

The original LIME framework perturbs input data by randomly altering feature values, leading to explanations that often lack domain-specific relevance or realism. To improve this, we redesigned the perturbation mechanism using **BioBERT**, a domain-specific language model, by converting input data into **textual descriptions**. This approach leverages BioBERT’s understanding of contextual nuances to generate meaningful perturbations that align with the underlying data distribution.

**Proposed Solution:**

1. **Input Conversion to Textual Description**:

• Each row of the tabular dataset is converted into a descriptive textual format that provides a natural language context for all the features.

• Example:

• Input Row:

Gender: Male, Age: 45, Household Size: 4, Income Poverty Ratio: 2.5, Body Mass Index: 28.0, Smoking Status: Former

A screenshot of a computer program

Description automatically generated

2. **Feature Masking**:

• Textual Description:

"A 45-year-old [MASK] living in a household of 4 people has an income-to-poverty ratio of 2.5. The individual has a body mass index of 28.0 and is a [MASK]."

Masked features allow the fine-tuned BioBERT model to predict plausible replacements based on the unmasked context.A black screen with white text

Description automatically generated

**Discovering Feature Relationships**:

• Using **CausalLearn** or similar tools, we analyzed the dataset to uncover causal relationships between features. For example:

• Income Poverty Ratio may causally influence Body Mass Index, which in turn might affect Health Status.

• These relationships are encoded in a causal graph, which helps prioritize feature dependencies.

A diagram of a network structure

Description automatically generatedA screenshot of a medical survey

Description automatically generated

**Prediction with BioBERT**:

• The masked textual description is passed to the fine-tuned BioBERT model, which predicts contextually relevant replacements for the masked features.

• Predicted replacements:

"Gender: Female" and "Smoking Status: Non-Smoker"

A screenshot of a computer

Description automatically generated

**Reconstruction of Perturbed Input**:

• The predicted values are converted back to the tabular format. For example:

Gender: Female, Age: 45, Household Size: 4, Income Poverty Ratio: 2.5, Body Mass Index: 28.0, Smoking Status: Non-Smoker

**Repeat for Multiple Perturbations**:

• This process is repeated for various combinations of masked features to generate a diverse set of perturbed samples. Each sample preserves the contextual relationships within the data.

6. **Model Predictions**:

• The perturbed samples are passed to the original machine learning model to obtain predictions for each perturbed instance.

7. **Local Model Training**:

• A local interpretable model (e.g., linear regression) is trained on the perturbed samples and their corresponding predictions. This local model approximates the global model’s behavior.

8. **Generating Explanations**:

• The local model provides feature importance scores, which explain the impact of each feature on the model’s predictions.

**Potential Customers**

1. **Data Scientists and Machine Learning Engineers**:

• **Why?**: They often rely on explainability tools to debug, improve, and validate machine learning models. The enhanced LIME approach offers context-aware and domain-relevant perturbations, providing more reliable and actionable insights.

• **Use Case**: Debugging a healthcare prediction model to understand why a specific patient’s risk score is high and how changes in variables like BMI or physical activity influence predictions.

2. **Regulatory and Compliance Teams**:

• **Why?**: Regulatory agencies and compliance officers require model explanations for audit trails, especially in sensitive domains like finance and healthcare.

• **Use Case**: Using the enhanced LIME to validate that a credit scoring model does not exhibit biases based on demographic variables like gender or income poverty ratio.

3. **Healthcare Professionals**:

• **Why?**: Doctors and medical researchers need interpretable AI models to support decision-making, especially when dealing with sensitive health data.

• **Use Case**: Explaining a diagnostic model’s prediction for chronic illnesses based on features like age, body mass index, and health status, making it easier for doctors to trust and act on the recommendations.

4. **Educators and Students in AI and Data Science**:

• **Why?**: Educators can use this project as a teaching tool to illustrate the importance of context-aware explanations, while students can learn by experimenting with realistic perturbations.

• **Use Case**: Incorporating the enhanced LIME framework in coursework for interpretability modules, allowing students to analyze how explanations improve with context-aware techniques.

5. **Business Analysts and Decision Makers**:

• **Why?**: Business stakeholders need simple and intuitive explanations of AI model outputs to make informed decisions.

• **Use Case**: Explaining why certain customer groups are more likely to churn or which factors most impact sales predictions in a retail setting.