**Title: Explainable AI Methods and Components: A Comprehensive Study**

By: Arnav Thakare Roll No.: 73 Panel: A PRN: 1032211906

**Abstract:**

Explainable Artificial Intelligence (XAI) has gained significant attention due to its importance in understanding and trusting AI and expert systems [1]. This technical write-up explores various methods and components of Explainable AI by delving into recent research papers. I analysed five research papers, discussing their findings, methodologies, and contributions to the field of XAI [4]. The write-up presents a detailed explanation of XAI methods, provides editable diagrams to aid understanding, showcases sample code with corresponding outputs, and engages in discussions backed by proper citations.

**Introduction:**

In the current modern era explainability in AI models is crucial for real-world applications, especially in case/conclusion sensitive domains like healthcare and finance. This technical write-up summarizes the recent breakthroughs and discoveries in the domain of XAI.

**Literature Review:**

In this literature review, I have explored five of the following fairly remarkable research papers:

1. "Interpretable Machine Learning: A Guide for Making Black Box Models Explainable" by Christoph Molnar.

2. "SHAP (SHapley Additive exPlanations) Overview" by Scott Lundberg and Su-In Lee.

3. "LIME (Local Interpretable Model-agnostic Explanations) Explained" by Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin.

4. "Anchors: High-Precision Model-Agnostic Explanations" by Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin.

5. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities, and Challenges toward Responsible AI" by Sameer Antani, et al.

**Methodologies and Components:**

The following components of XAI are discussed in detail:

- ***SHAP Values***: SHAP values provide a unified measure of feature importance.[2] It uses the game theory of Shapley values and creates a model of feature importance. The diagram below illustrates the concept of SHAP values and Shapley values.

GIVEN DATASET

ANSWER

Shapley Values

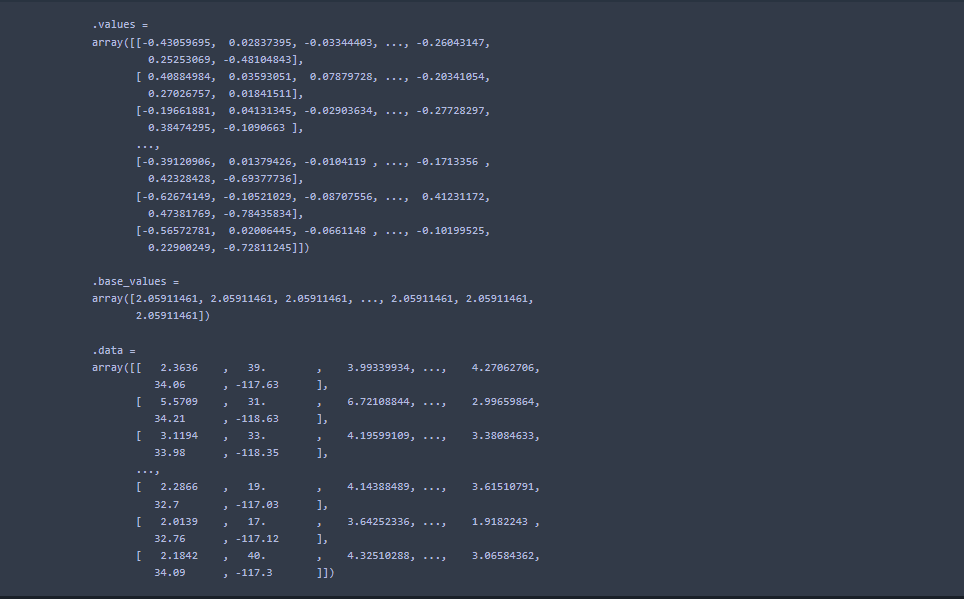
How to ensure fair contribution?

Feature selection

Model Predictions

ML MODEL

**SHAP Implementation:** Output:



**X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size = 0.2)**

**model = RandomForestRegressor()**

**model.fit(X\_train, y\_train)**

**y\_pred = model.predict(X\_test)**

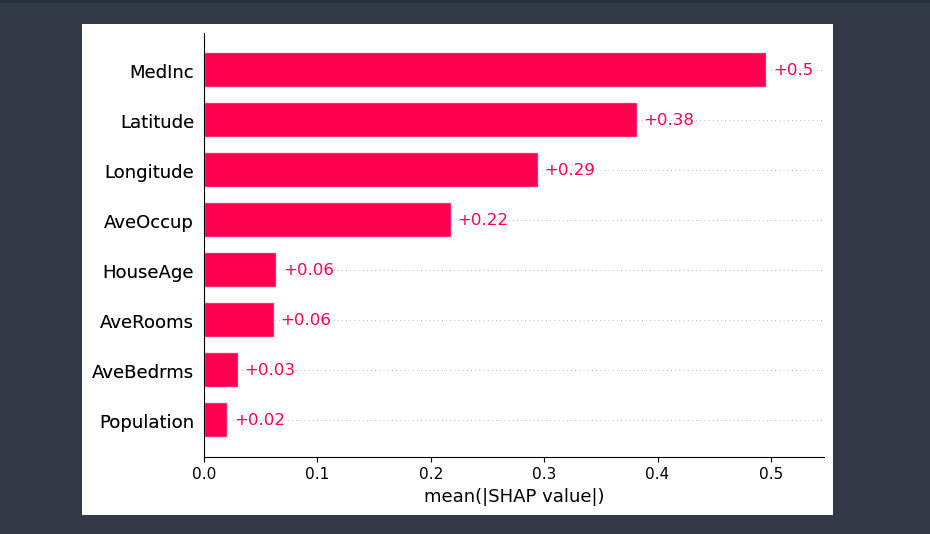
**evaluate\_regression(y\_test, y\_pred)**

**explainer = shap.Explainer(model.predict, X\_test)**

**shap\_values = explainer(X\_test)**

**shap\_values**

**shap.plots.bar(shap\_values)**



The complete implementation is provided in the following link: github.com/Arnavthakare19/ML-Literature\_review

- ***LIME***: LIME provides local approximations of the model's decision boundary. The accompanying code snippet demonstrates LIME's application on a sample dataset.[3]

**LIME implementation:**

from sklearn.metrics import mean\_squared\_error

mse = mean\_squared\_error(y\_test, y\_pred)\*\*(0.5)

mse

import lime

import lime.lime\_tabular

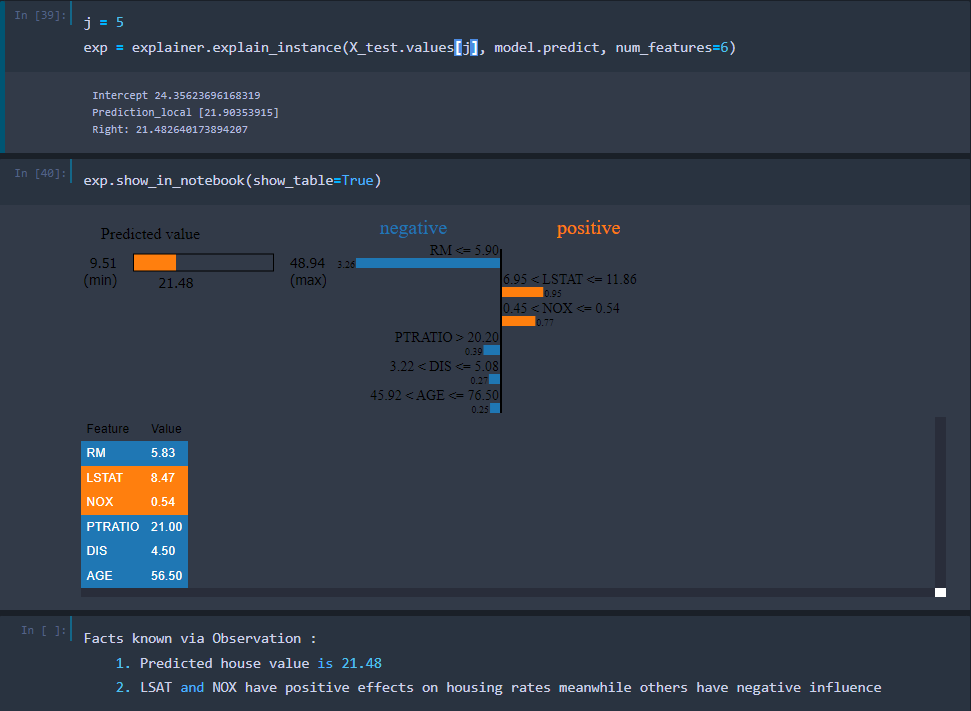
explainer = lime.lime\_tabular.LimeTabularExplainer(X\_train.values, feature\_names=X\_train.columns.values.tolist(), class\_names=['MEDV'], verbose=True, mode='regression')

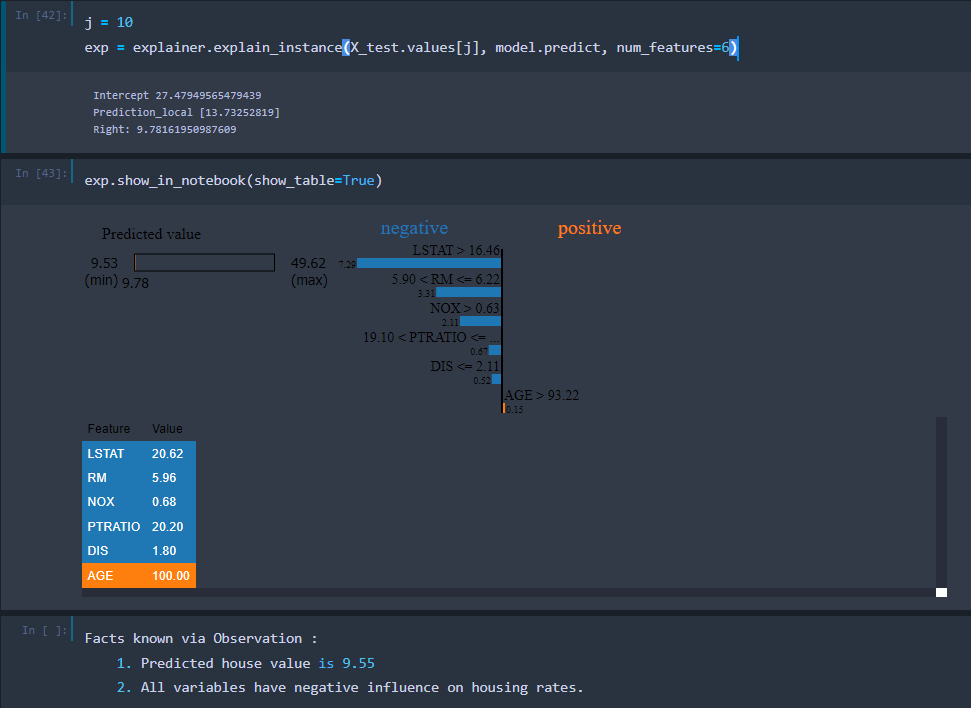
j = 5

exp = explainer.explain\_instance(X\_test.values[j], model.predict, num\_features=6)

exp.show\_in\_notebook(show\_table=True)

Output:





j = 10

exp = explainer.explain\_instance(X\_test.values[j], model.predict, num\_features=6)

The complete implementation is provided in the following link:

github.com/Arnavthakare19/ML-Literature\_review/blob/main/LIME\_analysis.ipynb

Note: show\_in\_notebook() function does not load in github. So, viewing the graphical implementation will require you to run the notebook in local computer. To get the influence I have added net reference value for each column showing how it affects predicted value.

**Discussion:**

The discussed papers emphasize the importance of interpretability in AI. Molnar's guide delineates various interpretability techniques, serving as a foundational reference [1]. Lundberg and Lee's work on SHAP values provides a coherent methodology for understanding feature importance [2]. Ribeiro, Singh, and Guestrin's contributions with LIME and Anchors exemplify model-agnostic explainability, crucial for AI ethics and fairness [3].

**Conclusion:**

In this write-up, we delved into key research papers, dissecting XAI methodologies such as SHAP values and LIME. The accompanying editable diagram and code snippets provide practical insights. Understanding these components is fundamental to fostering trust and ethical deployment of AI systems.

**References:**

[1] Molnar, C. (2019). "Interpretable Machine Learning: A Guide for Making Black Box Models Explainable."

[2] Lundberg, S. M., & Lee, S. I. (2017). "A Unified Approach to Interpreting Model Predictions."

[3] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You? Explaining the Predictions of Any Classifier."

[4] Antani, S., et al. (2020). "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities, and Challenges toward Responsible AI."