**Title: Explainable AI Methods and Components**

Roll No.: 73(Panel A) PRN: 1032211906

1. **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. It has gained significant attention due to its importance in understanding and trusting AI and expert systems [1]. The write-up presents a detailed explanation of various methods, provides editable diagrams to aid understanding, showcases sample code with corresponding outputs, and engages in discussions backed by proper citations.This technical write-up summarizes the recent breakthroughs and discoveries in the domain of XAI.

1. **Methodologies and Components:**

The following components of XAI are discussed in detail:

**2.1 *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. SHAP (SHapley Additive exPlanations) analysis in XIA (Explainable AI) assigns a value to each feature of a prediction, quantifying its impact. It computes the average contribution of a feature across all possible feature combinations, providing insights into individual feature importance within a model. The diagram below illustrates the concept of SHAP values and Shapley values.

GIVEN DATASET

How to ensure fair contribution?

ANSWER

Shapley Values

Feature selection

Figure 1. Shapley value game theory

Model Predictions

ML MODEL

**:**

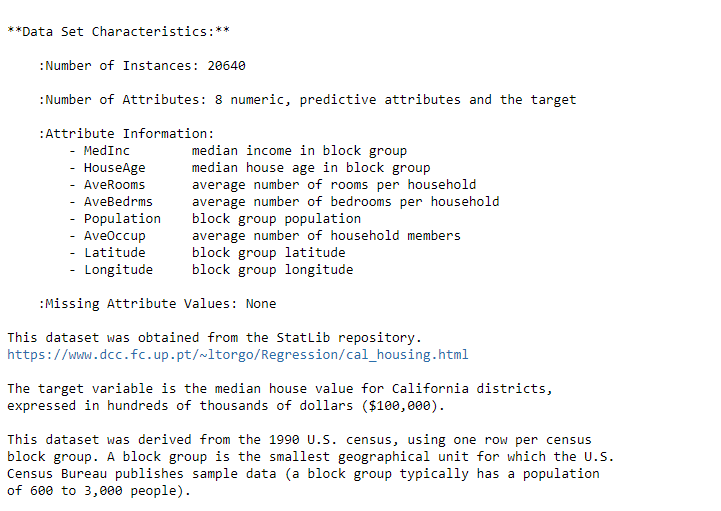
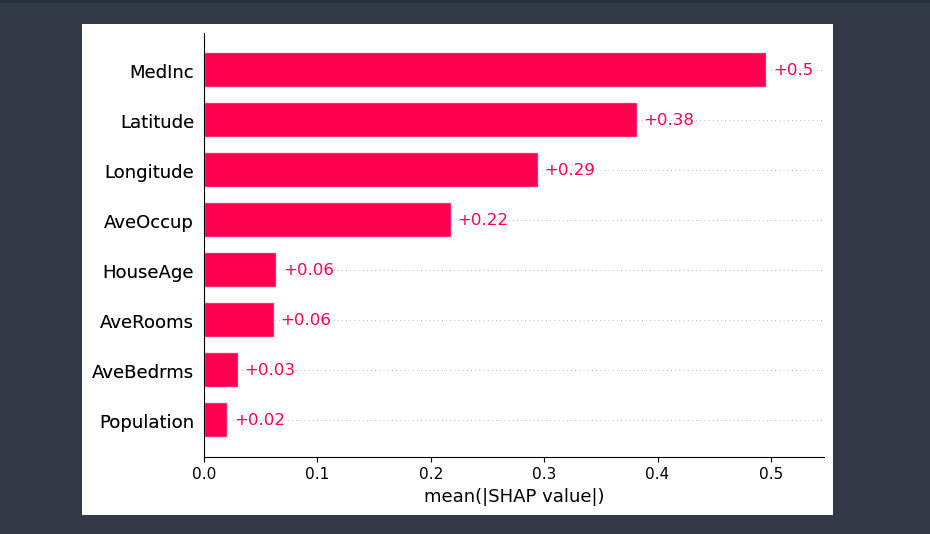


Image 2.1.1 Description of dataset used in the implementation

**SHAP Implementation**





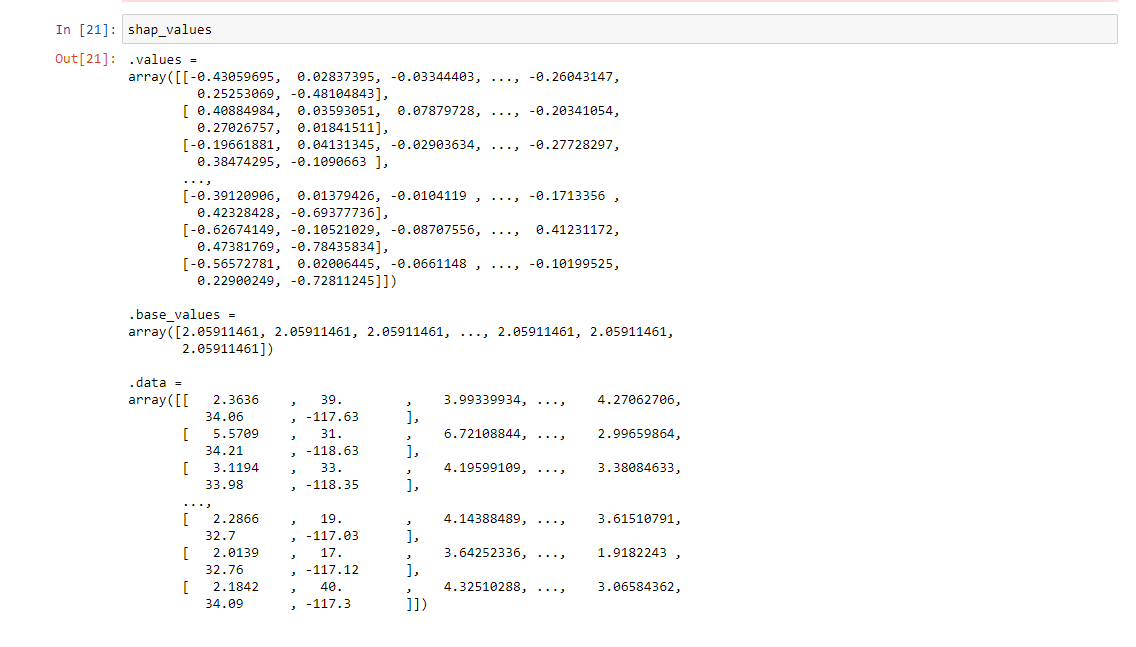


Image 2.1.2: SHAP values

**Explanation:**

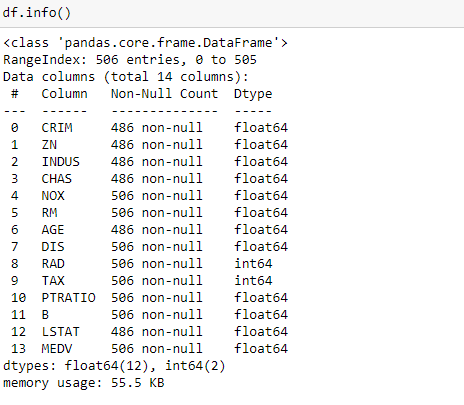
1. Imported the libraries and the dataset used.
2. Created the simple regression model.
3. Split the test train data and fit it into the model.
4. Evaluate the model.
5. Check feature importances
6. Fit the shap.Explainer using prediction model and train data.
7. Visualize the SHAP values and print the SHAP values

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

**2.2** ***LIME***: LIME (Local Interpretable Model-agnostic Explanations) in XIA creates local explanations for specific predictions.[3] It creates a simplified model around a pre-decided prediction point and observes how input features affect the output, offering insights into the model's behavior on a specific instance. LIME provides local approximations of the model's decision boundary. The accompanying code snippet demonstrates LIME's application on a sample dataset.

**LIME implementation:**

****

****

The different column data is described as follows:

1. 1 LSTAT - lower status of the population
2. 2 RM - average number of rooms per housing
3. 3 NOX - nitric oxides concentration (parts per 10 million)
4. 4 PTRATIO - pupil-teacher ratio by town
5. 5 DIS - weighted distances to five Boston employment centres
6. 6 AGE - proportion of owner-occupied units built prior to 1940

The target variable is MEDV which stands for **Median value of owner-occupied homes**.

Image 2.2.1 Description of dataset used

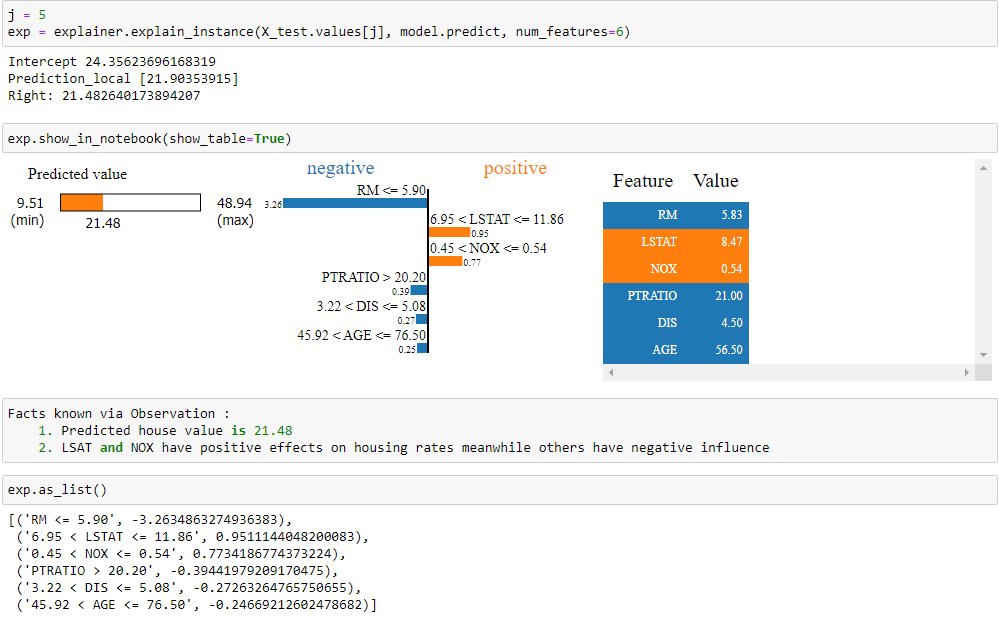
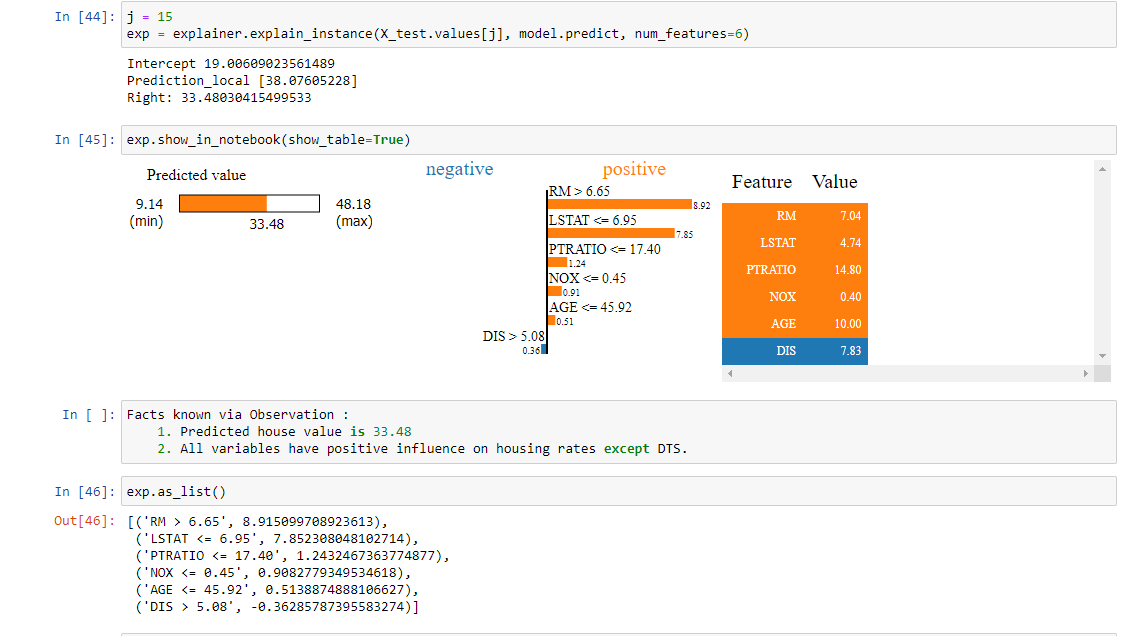


Figure 2.2.2 Corelation between house price to other factors in the dataset when j = 5



j = 10

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

Figure 2.2.3 Corelation between house price to other factors in the dataset when j = 15

**Explanation:**

1. Imported the libraries and the dataset used.
2. Split the data into train and test data.
3. Created the simple regression model with depth of 6 and 10 estimators for random forest model.
4. Fit it into the model.
5. Evaluate the model.
6. Check feature importances
7. Fit the lime. LimeTabularExplainer using prediction model and train data with the mode being ‘Regression’.
8. Use explainer.explain\_instance function while changing total number of instances in order to know which features affect the model the most and most optimized.

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.**

1. **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].

1. **Conclusion:**

In this write-up, we delved into key research papers, going through two currently used XAI methodologies such as SHAP values and LIME. The accompanying editable diagram and code snippets along with outputs provide practical implementation. Understanding these components is fundamental to fostering trust and ethical deployment of AI systems.

**5. References:**

[1] Molnar, C. (2019). "Interpretable Machine Learning: A Guide for Making Black Box Models Explainable." ISBN: 979-8411463330

[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] Alejandro Barredo Arrieta, Raja Chatila, Daniel Molina (2020). "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities, and Challenges toward Responsible AI."