Code here:
https://colab.research.google.com/drive/1qhopdNc9pR2LPEklFViLS0UnesQhwSjR?usp=sharing

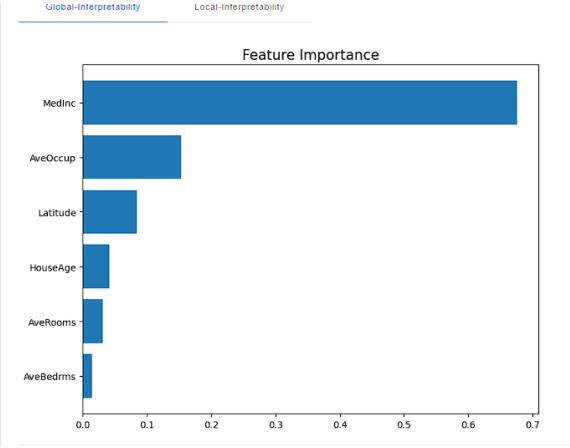
Model Explainability:
model explainability gives an explanation for why the model made a particular prediction or decision. Explainability requires not only that the model's decision-making process is transparent, but also that it is clear how the inputs and outputs of the model relate to the real world. An explainable model is not only transparent but also provides a clear rationale for its predictions or decisions

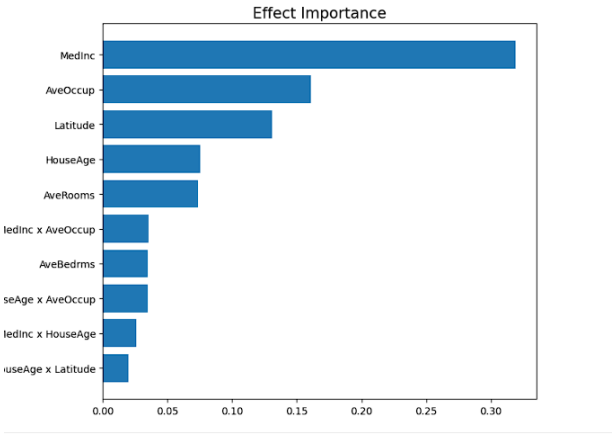
How do we check model explainability:
There are 2 methods:

1. Build an intrinsically interpretable model (white box)
2. Post hoc explanation of the model (black box)

What are the parts of post hoc analysis?

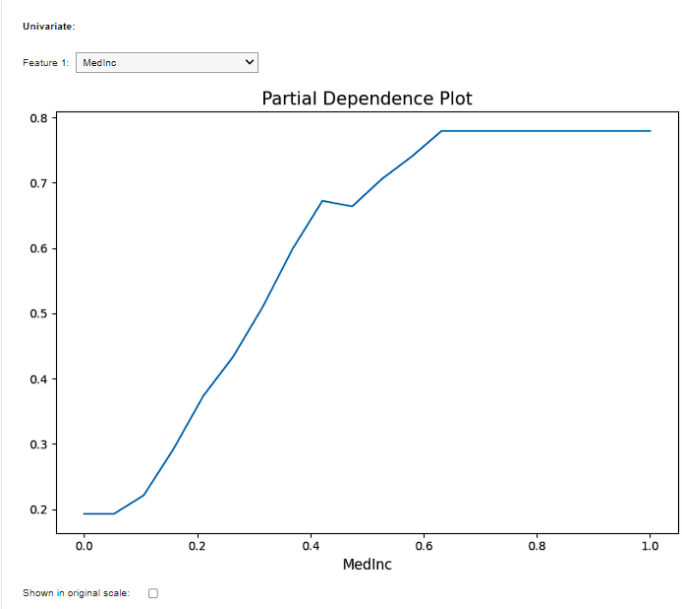
Global interpretation involves understanding how the model behaves across the entire input space. This may involve identifying the most important input features, exploring how the model responds to changes in these features, or generating visualizations to help understand the model's decision-making process. Global interpretation can provide insight into the overall behavior of the model, and help identify patterns or relationships that may be useful for further analysis or model improvement.

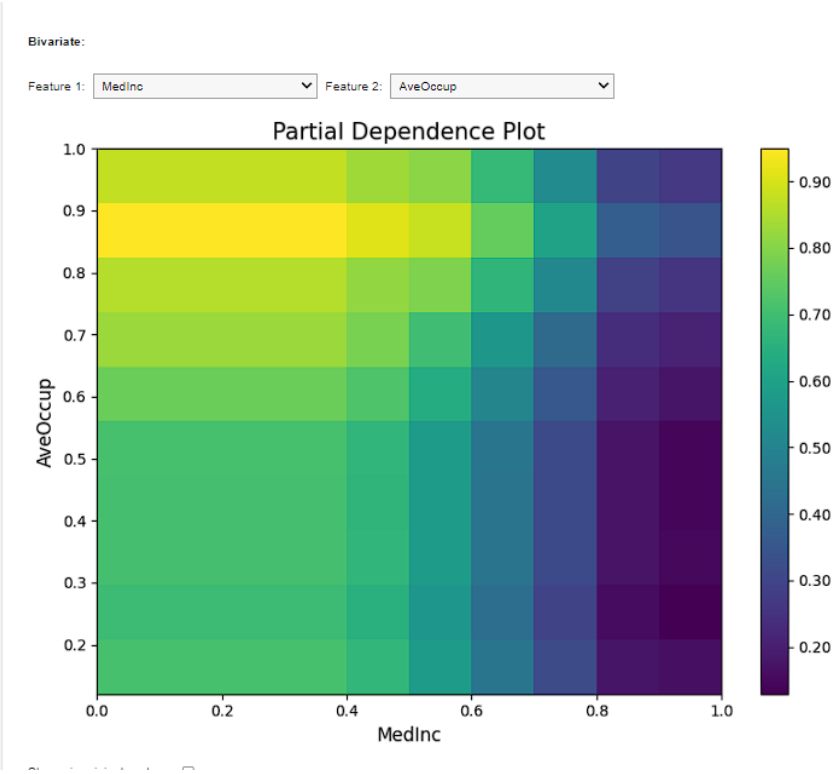




-PDP (Partial Dependency Plot)

A partial dependency plot displays the marginal effect of a feature on the predicted outcome. It allows us to observe how a feature and the target variable are related, while controlling for the effects of other variables. This can be useful in identifying non-linear relationships between features and the target variable, as well as identifying potential interactions between features.



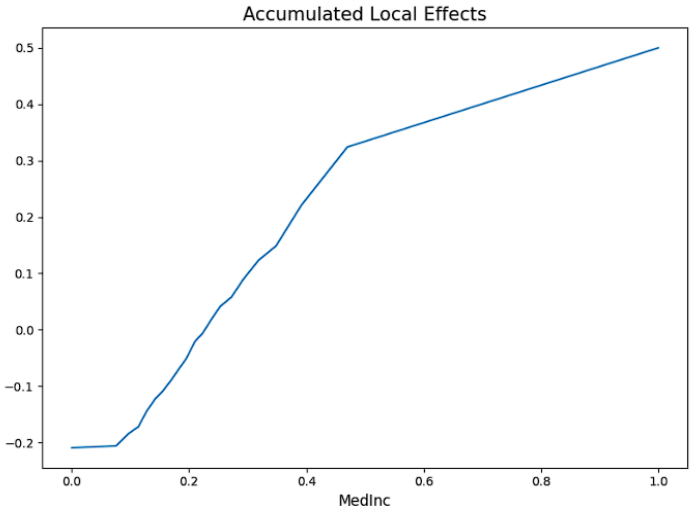


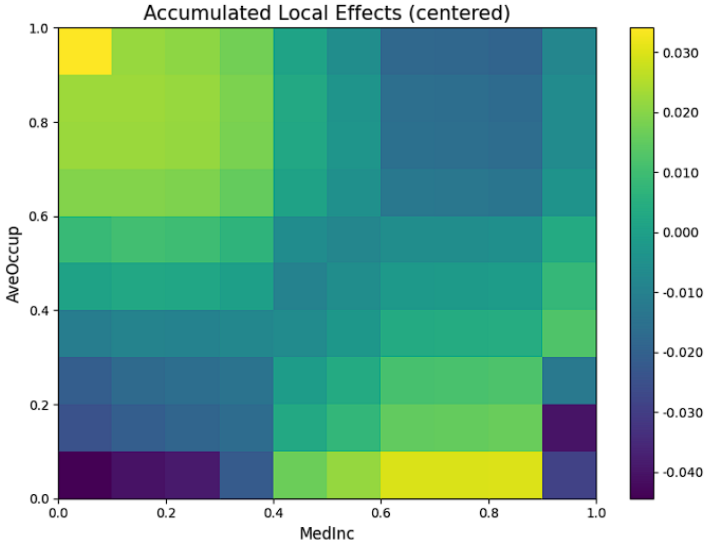
-ICE(Individual Conditional Expectation)

An Individual Conditional Expectation (ICE) is another graphical tool used for global interpretation. ICE plots visualize how the predicted outcome changes for each individual observation as a feature changes. It provides a more detailed view of how the model behaves for different instances and helps identify possible biases or areas of inconsistency in the model's decision-making process.

-ALE(Accumulated Local Effect)

ALE plots are similar to partial dependence plots (PDPs) in that they show the relationship between a feature and the predicted outcome in a model. However, unlike PDPs, ALE plots do not assume a linear relationship between the feature and the target variable. ALE plots show how the predicted outcome changes as a feature is varied, while controlling for the effects of all other features. ALE plots are created by first dividing the range of the feature into equally spaced intervals. For each interval, the average effect of the feature on the predicted outcome is calculated, while controlling for the effects of all other features. These average effects are then plotted against the midpoint of each interval, creating an ALE plot.

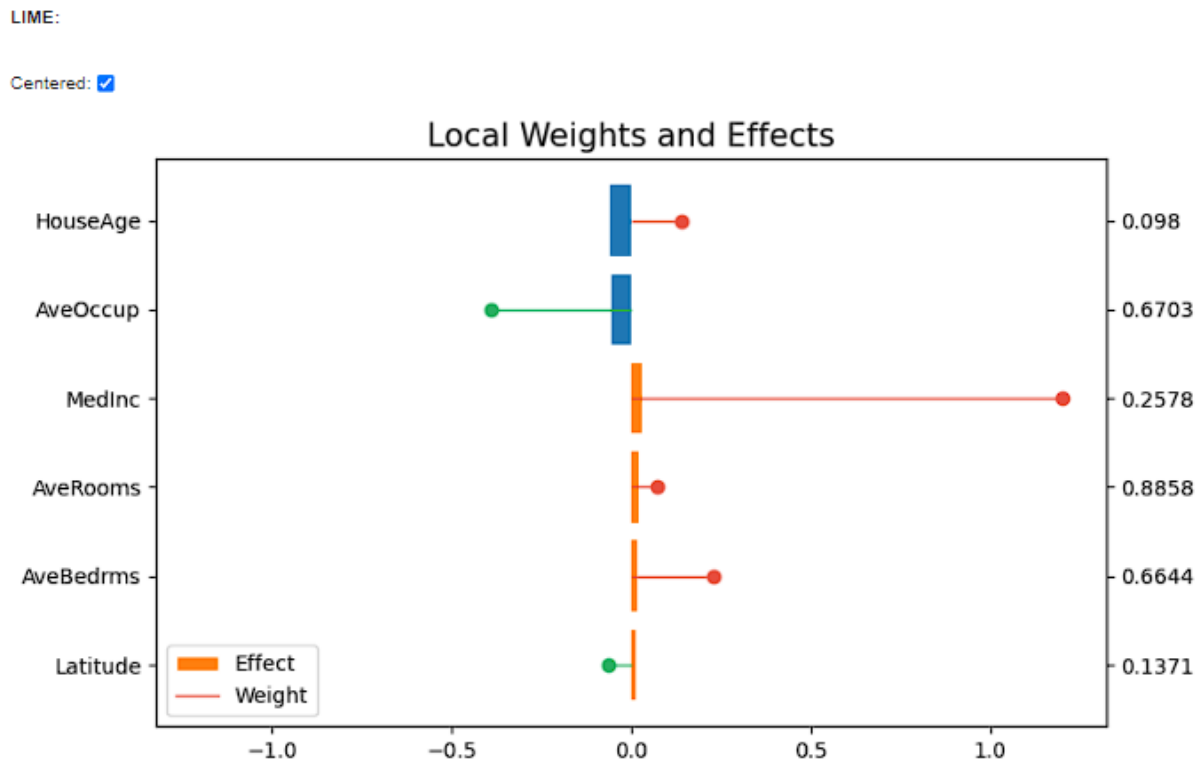

Accumulated Local Effects
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
MedInc


Accumulated Local Effects (centered)
AveOccup
MedInc
0.0
0.2
0.4
0.6
0.8
1.0
0.030
0.020
0.010
0.000
-0.010
-0.020
-0.030
-0.040

Local interpretation, on the other hand, involves understanding how the model behaves for a specific input or set of inputs. This may involve generating explanations or visualizations that help explain the model's prediction or decision for a particular instance, or exploring how the model responds to changes in the input features for that instance. Local interpretation can help identify cases where the model may be making unexpected or inconsistent predictions, and can help identify areas where the model may need to be improved.

-LIME (Local Interpretable Model-agnostic Explanations)

A method for explaining the predictions of a model on a local, instance-specific level. LIME works by generating a new dataset of perturbed samples around the instance of interest and training a local, interpretable model on this dataset. The local model is then used to explain the prediction of the original model by identifying the most important features for the prediction.

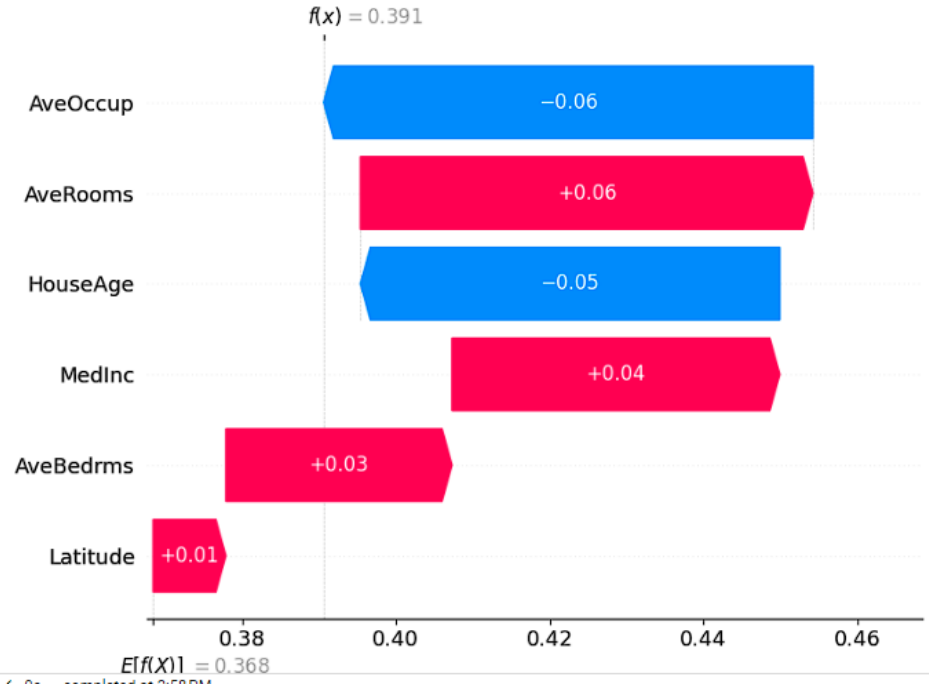


-SHAP (SHapley Additive exPlanations)

A method for explaining the predictions of a model on a local level. SHAP is based on game theory and computes the contribution of each feature to the difference between the predicted output and the expected output. It does this by considering all possible combinations of features and computing the marginal contribution of each feature to the prediction. This allows for a more accurate and nuanced explanation of the model's behavior.

SHAP:

Centered:



References:

"Machine Learning Interpretability and Explainability" by Christoph Molnar. Available online at https://christophm.github.io/interpretable-ml-book/

"A Unified Approach to Interpreting Model Predictions" by Scott Lundberg and Su-In Lee. Available online at https://arxiv.org/abs/1705.07874

"Explainable AI: Interpreting, Explaining and Visualizing Deep Learning" by Vaishak Belle. Available online at https://www.sciencedirect.com/science/article/pii/S2352711020300377

"Marginal Effects in Regression Analysis" by Thomas Carsey and Jeffrey Harden. Available online at https://methods.sagepub.com/reference/the-sage-encyc-of-political-science/n346.xml

"Partial Dependence Plots: A Tool for Exploring and Analyzing Machine Learning Models" by Jason Brownlee. Available online at https://machinelearningmastery.com/partial-dependence-plot-for-machine-learning/

"Local Interpretable Model-Agnostic Explanations (LIME)" by Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Available online at https://arxiv.org/abs/1602.04938

"A Unified Approach to Interpreting Model Predictions" by Scott Lundberg and Su-In Lee. Available online at https://arxiv.org/abs/1705.07874