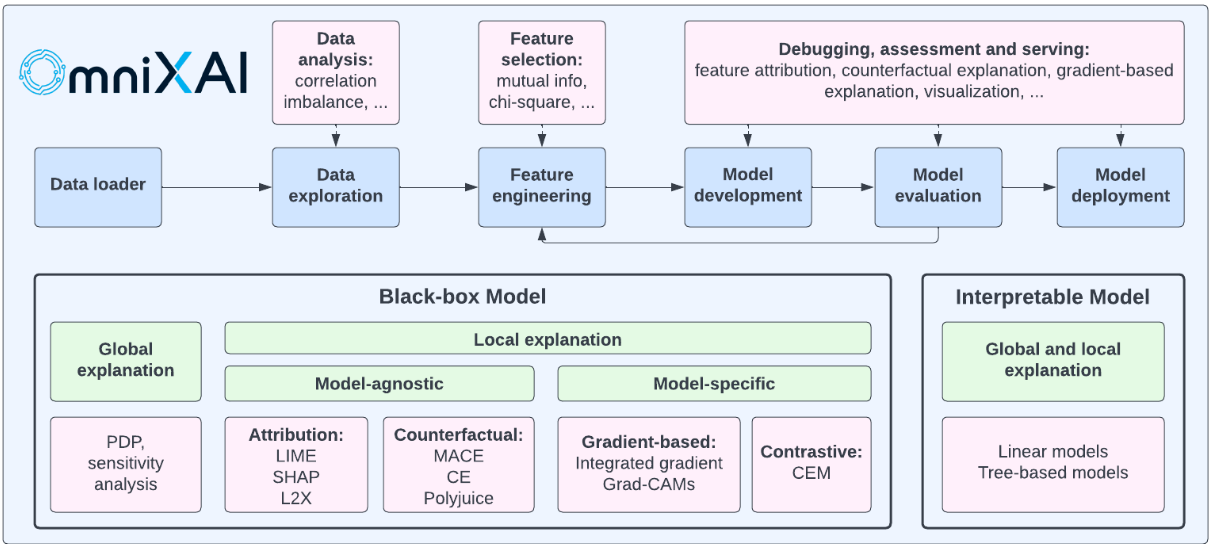
**Summary:**

With the increase in model complexity and the result of lack of transparency, model interpretability methods have become increasingly important. Model understanding is both an active area of research as well as an area of focus for practical applications across industries using machine learning.

Before we jump into the details, we should be aware that there is no single approach of explainability that works best. A summary of the approach is list as below:



However, this summary is not combined with our business directly, how to use the approach appropriately for a given use case is still confusion. By following the practice of our modeling process, which mainly in the tree-based machine learning model, we propose a framework which follow the below structure to answer the common modeling questions:

1. Understand the variables importance of model.

2. How variables contribute to the final prediction.

2.1 Understand variables interaction and its effect on final prediction.

2.2 Understand single variables effect on final prediction

3. Local explainability on specific data sample.

And in

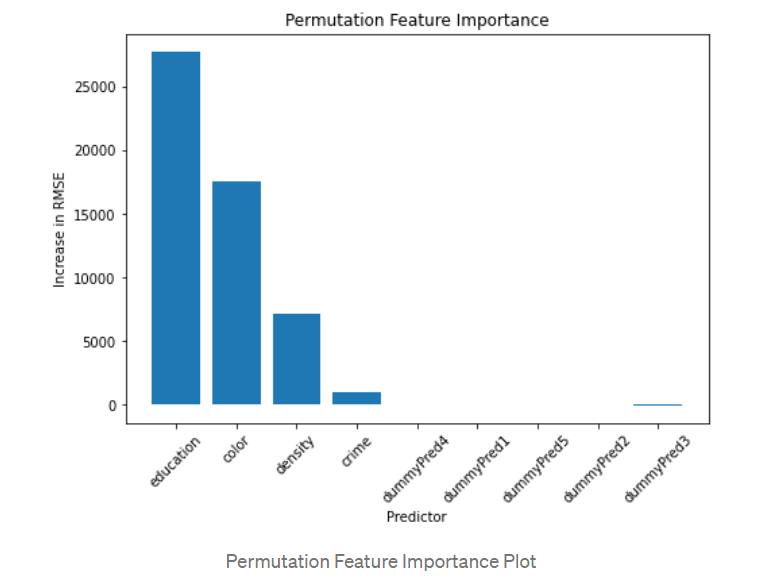
**Step 1: Variable Importance.**

**Common questions to be answered here:**

* *What are the top features in this model?*
* *What is the relative importance of feature X vs Y?*

Here we recommend **Permutation feature importance**: it measures the increase in the prediction error of the model after we permuted the feature’s values. This is model-agnostic and can overcome the feature importance’s bias of favor to numeric variables.

**How to use:** Scikit-learn provide an existing package. The permutation importance of a feature is calculated as follows. First, a baseline metric, defined by scoring, is evaluated on a (potentially different) dataset defined by the X. Next, a feature column from the validation set is permuted and the metric is evaluated again. The permutation importance is defined to be the difference between the baseline metric and metric from permutating the feature column. For classification, the score can be defined as roc\_auc. Below is an illustration to use RMSE (Root Mean Square Error) as score.



**Step 2: How variables contribute to the final prediction.**

How variables can contribute to the final prediction can separate into 2 scenarios:

1. The variable is independent, and the contribution is purely based on itself
2. The variable is correlated with other variables, so except for it owe effect, it’s interaction with other variables will also influence the final prediction.

To explore the different scenarios, we explore 2 parts of evaluation. First, we try to understand whether the variables’ interaction will influence the final prediction, then we look into single variables to understand how it can influence the final prediction.

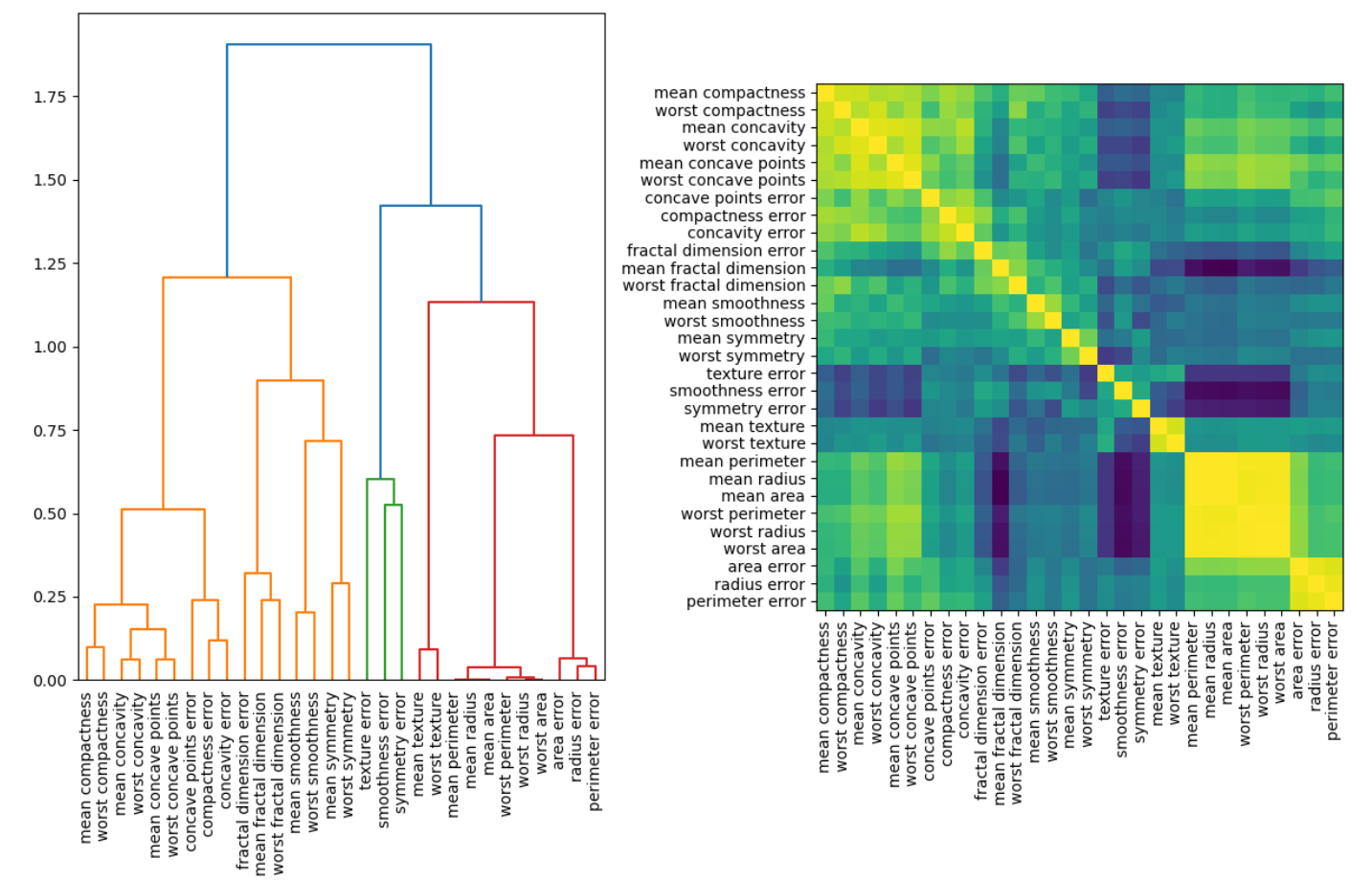
**Step2.1 Variable’s interaction and its effect on final prediction.**

**Common questions to be answered here:**

* *Does that model contain two (or more) highly correlated both features?*
* *Is the correlated variables interaction have a strong effect on the final prediction?*
* *How will the correlated variables interaction contribute to the final prediction?*

Here we propose 3 approaches combined to evaluate:

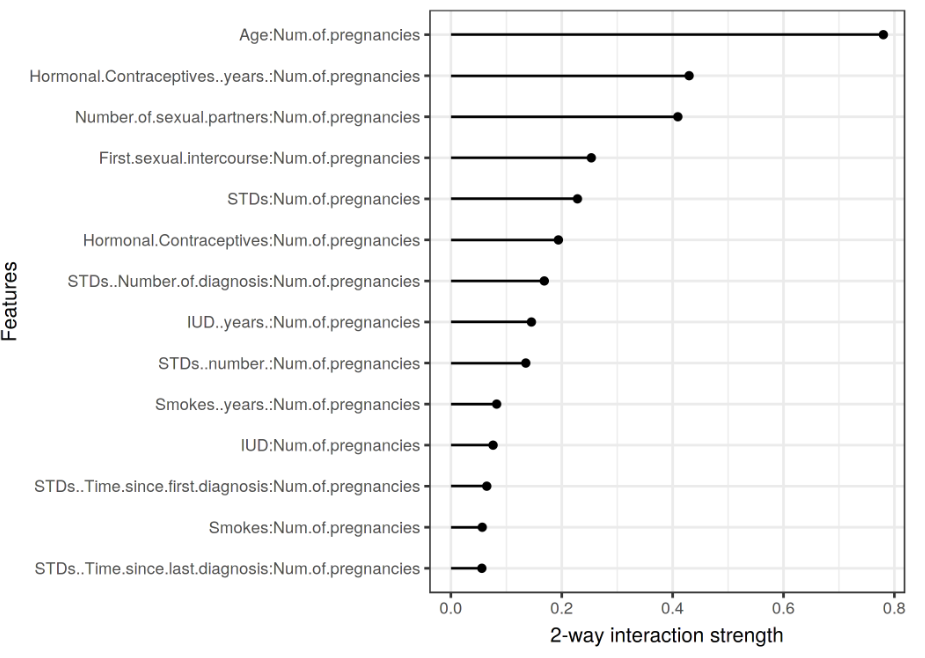
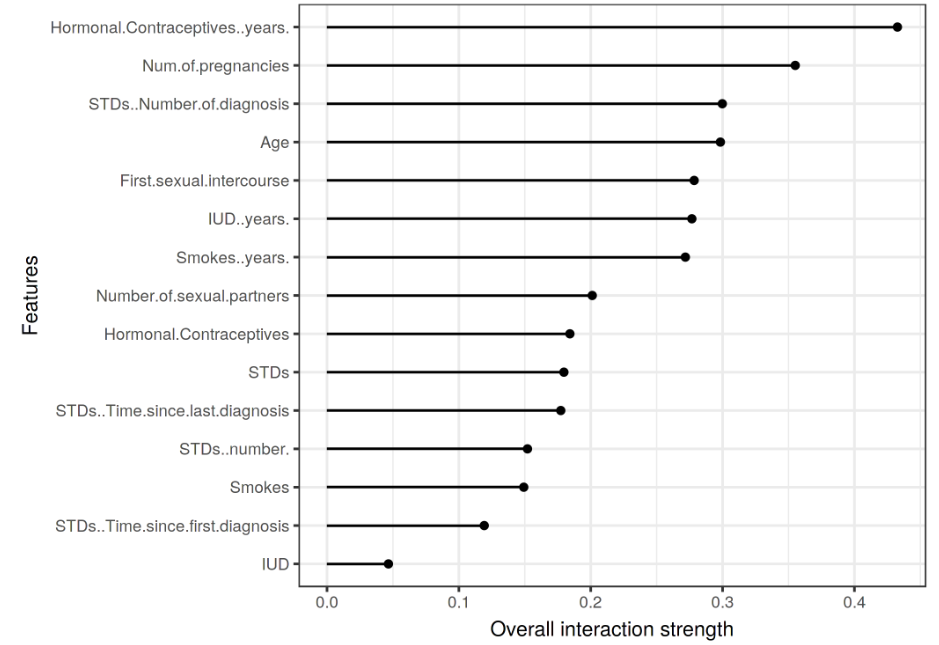
1. **Variables Cluster.** Based on different cutoff of variables correlation, variables cluster into different group. This can give us a view on how variables correlated each other. Usually, correlation large than 0.5 will consider as a strong correlation.



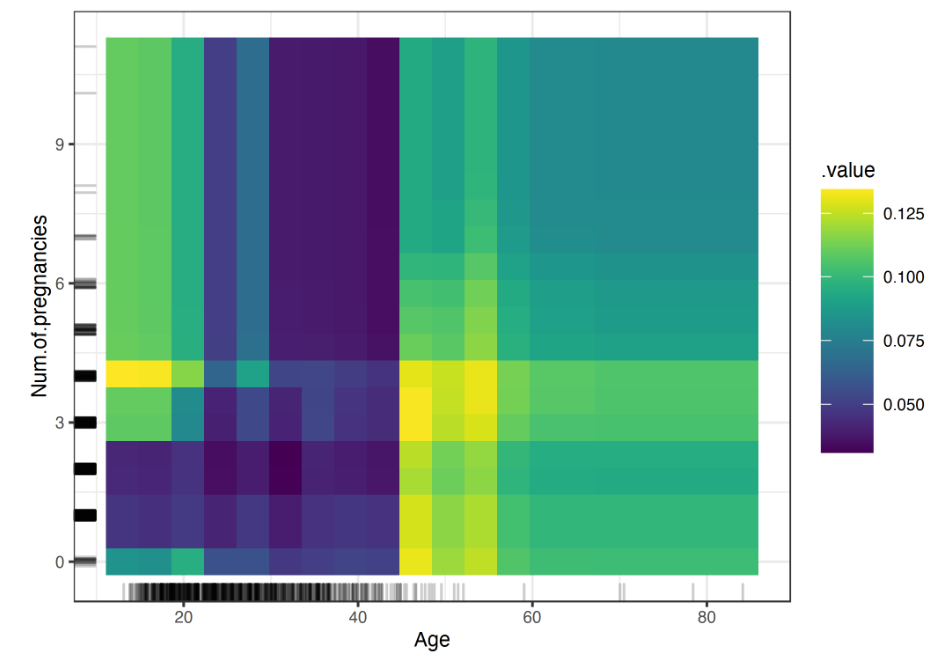
**Variables Clusters:**

1. **Friedman’s H-statistic:** Then for the high correlated variables, we can check the value of H-statistic on whether this correlation contribute a lot in the final prediction. If H-statistic is10%, then it means 10% of variance is explained by this feature interaction with other variables. There don’t have rule of thumbs for H-statistic, but general 10%-20% is relatively low and above 30% then we may should pay attention.

**Friedman’s H-statistic**



1. **2D PDP plots.**  Show the combined effect of the correlated features. (Kindly notice if we are only interested in the interaction effect, we should look at the second-order effects on ALE 2D plots, because the PDP plots is focus on total effect which already mixes the main effects into the plot)



**2D PDP plots:**

**How to use:** To combine these 3 approaches into 1, we propose a scenario here. For example, if we find variables A is very important meanwhile is highly correlated with variables B. Meanwhile there is a high H-statistic between A and B, then we should be careful to delete variables B. As it seems lots of A’s prediction power is come from the interaction with variables B.

**Step2.2 Understand how single variables correlated with target**

**Common questions to be answered here:**

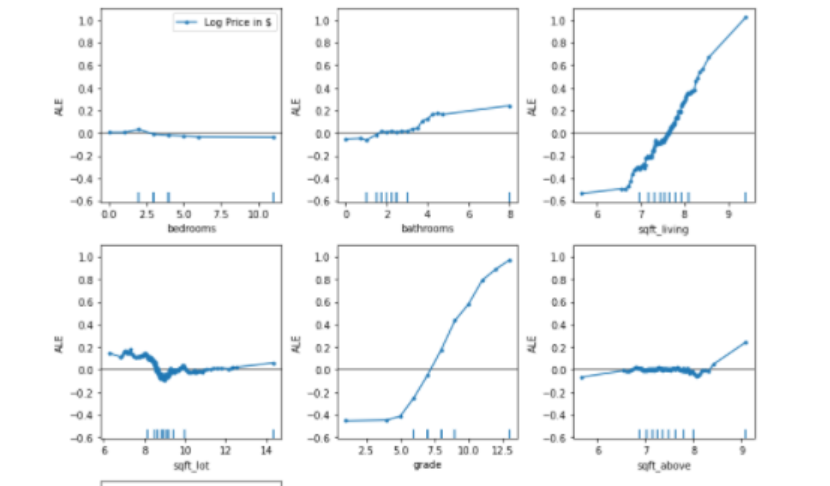
* *What is the relationship between feature X and target?*

The common technique is Partial Dependence Plot(PDP plot) and Accumulated local effects (ALE) plot.

The partial dependence plot (short PDP or PD plot) shows the average marginal effect one or two features have on the predicted outcome of a machine learning model. A partial dependence plot can show whether the relationship between the target and a feature is linear, monotonic or more complex.

Here we recommend **ALE** over PDP. ALE describe how features influence the prediction of a machine learning model on average. But It’s a faster and unbiased alternative to partial dependence plots (PDPs).

* The value of the ALE can be interpreted as the main effect of the feature at a certain value compared to the average prediction of the data.
* For example, an ALE estimate of -2 at xj=3xj=3 means that when the j-th feature has value 3, then the prediction is lower by 2 compared to the average prediction.



**Accumulated Local Effects (ALE) – Feature Effects Global Interpretability**

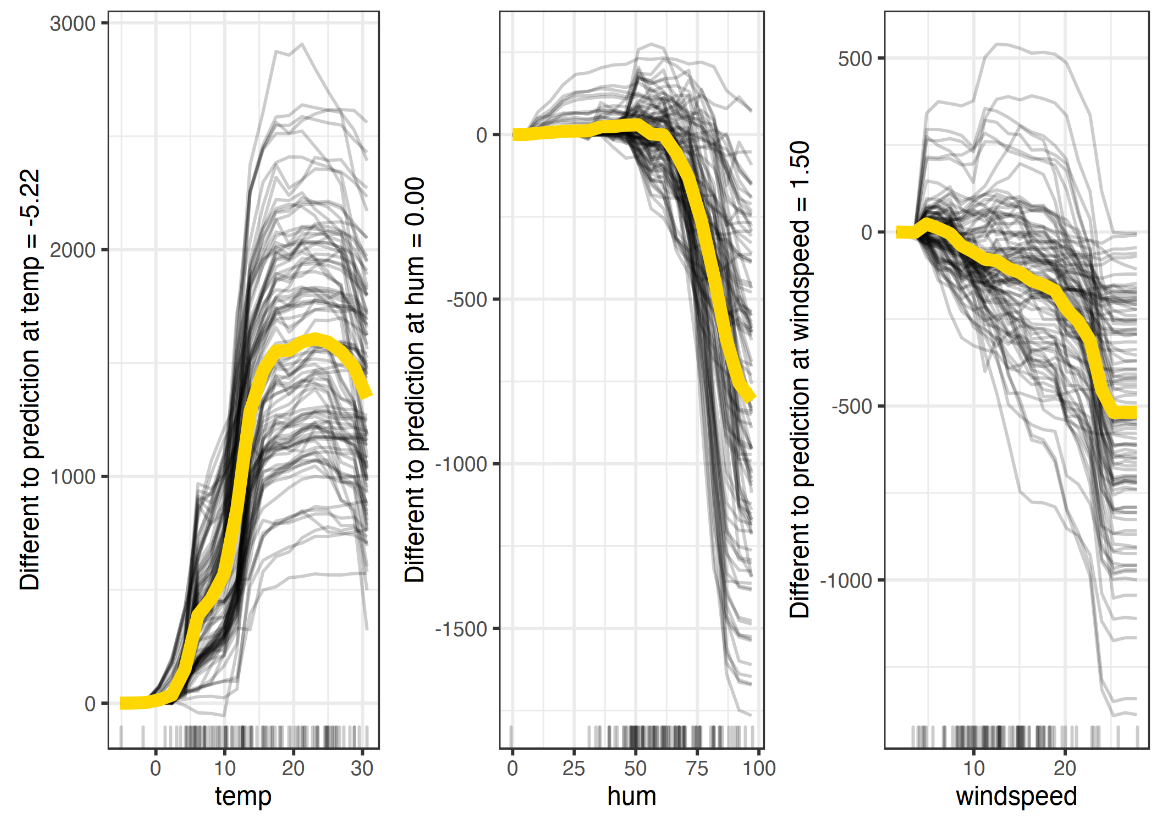
**Step3 – Local explainability on specific data sample**

**3.1 Aggerated local effect.**

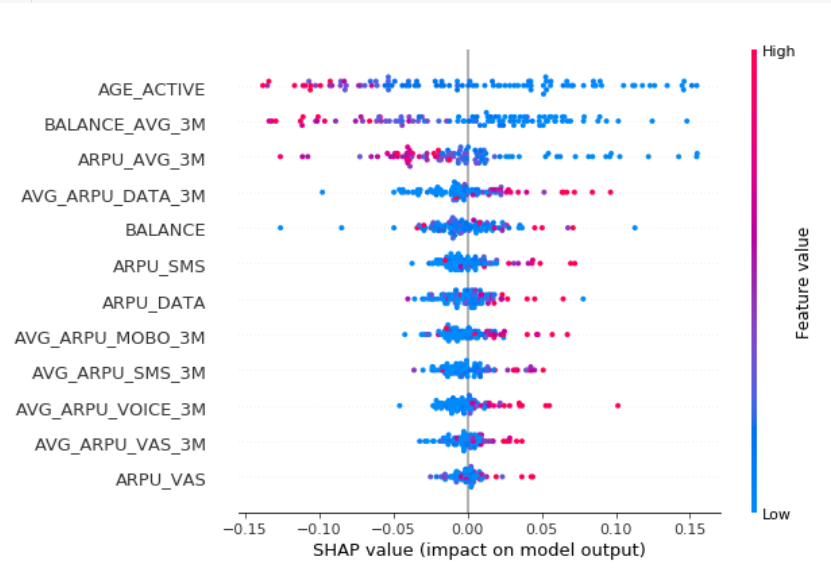
**Common questions to be answered:**

* *Does the feature behave consistently across the whole sample?*
* *Is there any outlier in the data?*

**Individual Conditional Expectation (ICE):** display one line per instance that shows how the instance’s prediction changes when a feature changes.

****

**Shap effect on each instance:** from the graph below, each point represents one instance. For a variables with strong predictive power, we should expect the point will spend wider, as it can separate the target well. And for the variables concentrated near 0, the predictive power will be relative low.

****

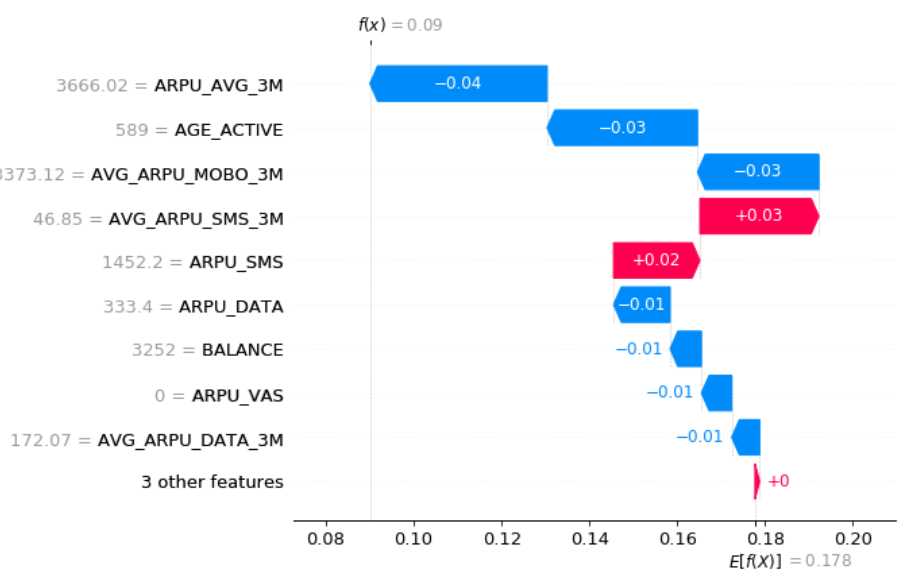
**3.2 Local explanation:**

**3.2.1 Shap Value:**

The goal of SHAP is to explain the prediction of an instance x by computing the contribution of each feature to the prediction. The SHAP explanation method computes Shapley values from coalitional game theory. The feature values of a data instance act as players in a coalition. Shapley values tell us how to fairly distribute the “payout” (= the prediction) among the features.

**Common questions to be answered:**

* *Why was this record classified as 0/1?*
* *What are the top reasons for assigning a higher (or lower) score for the record?*
* *Why is record X ranked higher than record Y?*

**

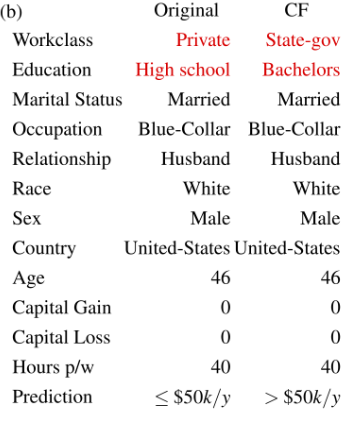
**3.2.2 counterfactual explanation:**

**A counterfactual explanation of a prediction describes the smallest change to the feature values that changes the prediction to a predefined output.**

**Common questions to be answered:**

* *What features does customer / record X need to improve in order to improve their score by Y points?*
* *What is the sensitivity of feature X around this record (change in score given an x% change in feature value)?*

**Example:** below showcase when the applier change the work class and education,the prediction of income will change accordingly.



**4. A summary on existing open-source resource:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Data Type** | **Method** | **OmniXAI** | **InterpretML** | **AIX360** | **Eli5** | **Captum** | **Alibi** | **explainX** |
| Tabular | **Github Stars** | **0.28k** | **4.8k** | **1.1k** | **2.5k** | **3.3k** | **1.7k** | **0.31k** |
| Permutation Importance |  |  |  | ✅ |  |  |  |
| Variables Cluster |  |  |  |  |  |  |  |
| Friedman’s H-statistic |  |  |  |  |  |  |  |
| 2D PDP plots |  |  |  |  |  |  |  |
| PDP | ✅ | ✅ |  |  |  |  | ✅ |
| ALE |  |  |  |  |  | ✅ |  |
| Individual Conditional Expectation (ICE) |  |  |  |  |  |  |  |
| Shap effect on each instance |  |  |  |  |  |  |  |
| LIME | ✅ | ✅ | ✅ | ✅ | ✅ |  |  |
| Shap | ✅ | ✅ | ✅ |  | ✅ | ✅ | ✅ |
| Counterfactual | ✅ |  | ✅ |  |  | ✅ | ✅ |
| **Note** | |  | Also have explain for glassbox model |  |  | Model Interpretability for PyTorch |  |  |