This document explains the methodology implemented to compute ICE based and Shapley value based individual prediction explanations in Dataiku DSS.

In both cases, if the user wants *N* explanations, then the explanations are computed for the 10*N first features ranked by global feature importance and only the *N* explanations with the largest absolute values are returned to the user. If the model does not provide feature importances, they are computed by training a random forest.

ICE based explanation methodology

Let us define the prediction y and the prediction function f: $y = f(X)$ with $X = \left[x_i\right]_{i \in [1,I]}$

Let us define a distribution of X represented by a dataset containing the samples $X^j, j \in [1,J]$. The samples may be weighted. The weight associated to each sample is $w^j > 0$.

The individual prediction explanation for the feature i of sample X is defined as:

$$\varphi_i(f(.),X) = y - \bar{y} \text{ with } \bar{y} = \sum_j w^j f\left(X^j_{frankenstein}\right) / \sum_j w^j$$

Where the 'Frankensteins' are defined such that they are identical to X except for the feature i for which the explanation is computed: they have the feature value of X^j instead of the one of X:

$$X^j_{frankenstein} = \left[x_1, ..., x_{i-1}, x^j_i, x_{i+1}, ..., x_I\right]$$

The distribution dataset is the test dataset in case split testing was selected and the full dataset if cross-testing was. To simplify and speed up the computation:

- When the feature i is numerical:
  - if it has more than 10 modalities, bins based on the weighted deciles are made and the average $\bar{y}$ is computed as the 10 bins weighted median.
  - else it is treated as a categorical variable.
- When the feature is categorical:
  - if the 10 more frequent modalities account for more than 90% of the total weight, then the average $\bar{y}$ is computed on these 10 modalities, weighted proportionally to their weighted frequency.
  - else, it is treated as a text feature (see below).
- When the feature is textual, a background of 25 samples is randomly drawn in the distribution dataset (only once for all features). The average $\bar{y}$ is computed on these 25 optionally weighted samples, after aggregation of those with the same modality.

Note: 'ICE' means 'Individual Conditional Expectation'. For continuous and ordinal features, the ICE plot is a plot of the predictions obtained Frankensteins based on a representative set of feature values. $\bar{y}$ is a weighted average of this predictions.

Shapley value based explanation methodology

Based on the notations for the ICE methodology, the individual prediction explanation for the feature i of sample X is now defined as the average over $j \in [1, 100]$ background samples of the Shapley value estimation for the background sample X^j :

$$\varphi_i(f(.), X) = \sum_j w^j \varphi_i(f(.), X, X^j)$$

This Shapley value is estimated as the average of the impact of switching from the value of feature i in X^j to the value of feature i in X while random feature values have already been switched:

$$\varphi_i(f(.), X, X^j) = \frac{1}{K} \sum_k \varphi_i^k(f(.), X, X^j)$$

Mathematically speaking, $\varphi_i^k(f(.), X, X^j)$ can be defined in this way:

- Let's define a permutation k of $[1, I]$: $\tau_k(u) = v$
- Let's define the two 'Frankensteins' that are used to evaluate the impact of the feature of interest i while using the permutation k. "Before" the feature of interest (i.e. for features with indices below the indice of the feature of interest in the current permutation), the Frankensteins are identical to the background sample X^j, while "after" the feature i, they are identical to the sample of interest X. They differ only for the value of feature i.
  - 'Start Frankenstein' $X^k_{start\ frank.}$ definition: $x_u^{k,s} = x_u^j$ if $\tau_k(u) \leq \tau_k(i)$ and $x_u^{k,s} = x_u^i$ if $\tau_k(u) > \tau_k(i)$
  - 'End Frankenstein' $X^k_{end\ frank.}$ definition: $x_u^{k,e} = x_u^j$ if $\tau_k(u) < \tau_k(i)$ and $x_u^{k,e} = x_u^i$ if $\tau_k(u) \geq \tau_k(i)$
- The impact of the feature given the permutation is the difference of "Frankensteins" predictions:

$$\varphi_i^k(f(.), X, X^j) = f(X^k_{end\ frank.}) - f(X^k_{start\ frank.})$$

Note: if only the permutation in which i is the first feature to be switched is used, then Shapley value based explanation methodology is equivalent to ICE based methodology. Indeed, in this case, the 'end Frankenstein' is always equal to the sample of interest. It means that ICE based methodology can be seen as a simplification of Shapley value based methodology.