Methods for Explaining Individual Predictions

Dataiku DSS
www.dataiku.com

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

Introduction

In both cases, if the user wants N explanations, then the explanations are computed for the $5 \times N$ 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 surrogate model and using its feature importances.

ICE based explanation methodology

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

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:

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

Where the 'Frankensteins' are defined as 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:

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

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 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 most 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 frequency.
  - else, it is treated as a text feature.
- When the feature is textual: the average $\bar{y}$ is computed on the 25 most frequent modalities, weighted proportionally to their frequency.

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 these predictions.

Shapley value based explanation methodology

The method stems from the ideas outlined in the SHAP (SHapley Additive exPlanations) framework [1].

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 averaged 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 us define a permutation k of $[1, I]$: $\tau_{k(u)} = v$
- Let us 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' $X^k_{\text{start frank.}}$ definition:

  $$x_u^{k,s} = \begin{cases} x_u^j \text{ if } \tau_{k(u)} \leq \tau_{k(i)} \\ x_u^i \text{ if } \tau_{k(u)} > \tau_{k(i)} \end{cases}$$

  'End Frankenstein' $X^k_{\text{end frank.}}$ definition:

  $$x_u^{k,e} = \begin{cases} x_u^j \text{ if } \tau_{k(u)} < \tau_{k(i)} \\ x_u^i \text{ if } \tau_{k(u)} \geq \tau_{k(i)} \end{cases}$$

- The impact of the feature given the permutation is the difference between the two "Frankensteins" predictions:

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

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 the ICE based methodology can be seen as a simplification of the Shapley value based methodology.

Illustration

The idea underlying Shapley value estimation is to **randomly swap columns** of the row to explain with a random sample of the training data, called the **background sample**.

Given the following row for which to explain the prediction:

Paris	38	100	French

And the following draw of background sample from the train set:

Berlin	40	200	Japanese
Rio	50	300	Nigerian
London	60	350	Belgian
Lisbon	20	1000	Italian
Dubai	35	800	Peruvian

Draw some random **permutations**, each row containing 1 to I (number of columns), shuffled :

1	2	3	4
4	2	1	3
3	2	4	1
2	1	3	4
4	1	2	3

To estimate the impact of **feature #3**, build two arrays of J rows (number of rows in the background sample), such as, for each row:

'Start Frankenstein':

Columns up to **and including 3** are taken from the background sample, otherwise from the row to explain.

1	2	3	4
4	2	1	3
3	2	4	1
2	1	3	4
4	1	2	3

Berlin	40	200	French
Rio	50	300	Nigerian
Paris	38	350	French
Lisbon	20	1000	French
Dubai	35	800	Peruvian

'End Frankenstein':

Columns up to **not including 3** are taken from the background sample, otherwise from the row to explain.

1	2	3	4
4	2	1	3
3	2	4	1
2	1	3	4
4	1	2	3

Berlin	40	100	French
Rio	50	100	Nigerian
Paris	38	100	French
Lisbon	20	100	French
Dubai	35	100	Peruvian

The impact of feature #3 is given by averaging the differences in prediction (value for regression, log-odds for classification) between End and Start Frankensteins.

Bibliography

[1] Scott Lundberg and Su-In Lee, "A Unified Approach to Interpreting Model Predictions." 2017. doi: 10.48550/arXiv.1705.07874.