Explaining individual predictions when features are dependent: More accurate approximations to Shapley values

Kjersti Aas   Martin Jullum   Anders Løland

Norwegian Computing Center

Summary

We want to explain individual predictions from machine models by learning simple, interpretable explanations.

- Shapley values is a game theoretic concept that can be used for this purpose.
- The Shapley value framework has a series of desirable theoretical properties, and can in principle handle any predictive model.
- Kernel SHAP is a computationally efficient approximation to Shapley values in higher dimensions.
- Like several other existing methods, this approach assumes that the features are independent. Since Shapley values currently suffer when features are correlated, the explanations may be very misleading.
- **We extend the Kernel SHAP method to handle dependent features.**

Introduction

The exact computation of Shapley values becomes intractable for more than, say, ten features. This has led to approximations like the Shapley Sampling Values [1] and Kernel SHAP [2]. The latter requires less computational power to obtain a similar approximation accuracy. Hence, we focus on improving the Kernel SHAP method to account for dependence.

In observational studies and machine learning problems, it is very rare that the features are statistically independent, meaning that the Shapley value methods suffers from inclusion of predictions based on unrealistic data instances when features are correlated. This is the case even if a simple linear model is used.

Our approach is implemented in the R-package `shapr` [3], which is available on `cran.r-project.org`.

The Shapley values setting

A training set $\{y^i, \boldsymbol{x}^i\}_{i=1,\dots,n_{\text{train}}}$ of size n_{train} has been used to train a predictive model $f(\boldsymbol{x})$ to resemble a response value y. We want to explain the prediction from the model $f(\boldsymbol{x}^*)$, for a specific feature vector $\boldsymbol{x} = \boldsymbol{x}^*$. The prediction $f(\boldsymbol{x}^*)$ is decomposed as

$$f(\boldsymbol{x}^*) = \phi_0 + \sum_{j=1}^{M} \phi_j^*,$$

where $\phi_0 = v(\emptyset)$ and ϕ_j^* is the ϕ_j for the prediction $\boldsymbol{x} = \boldsymbol{x}^*$. The Shapley values explain the difference between the prediction and the global average prediction. Here,

$$\phi_j = \sum_{\mathcal{S} \subseteq \mathcal{M} \setminus \{j\}} \frac{|\mathcal{S}|!(M - |\mathcal{S}| - 1)!}{M!} (v(\mathcal{S} \cup \{j\}) - v(\mathcal{S})),$$
$$j = 1, \dots, M.$$

The key ingredient here is the contribution function $v(\mathcal{S}) = \mathrm{E}[f(\boldsymbol{x}) | \boldsymbol{x}_{\mathcal{S}} = \boldsymbol{x}^*_{\mathcal{S}}]$ for a certain subset $\mathcal{S}$. This function should resemble the value of $f(\boldsymbol{x}^*)$ when we only know the value of the subset $\mathcal{S}$ of these features.

A brief overview of Kernel SHAP

The Kernel SHAP method [2] tries to estimate $v(\mathcal{S})$ in practical situations and consists of two parts:

① A clever computationally tractable approximation for computing the Shapley values – approximated least squares

② A simple Monte Carlo integration method for estimating $v(\mathcal{S})$ – assuming $p(\boldsymbol{x}_{\bar{\mathcal{S}}} | \boldsymbol{x}_{\mathcal{S}} = \boldsymbol{x}^*_{\mathcal{S}}) \approx p(\boldsymbol{x}_{\bar{\mathcal{S}}})$ [independence!]

Linearity and independence is simple

We show that *if the predictive model is a linear regression model* $f(\boldsymbol{x}) = \beta_0 + \sum_{j=1}^M \beta_j x_j$, where all features $x_j, j = 1, \dots, M$ are *independent*, then the Shapley values take the simple form: $\phi_0 = \beta_0 + \sum_{j=1}^M \beta_j E[x_j]$ and $\phi_j = \beta_j\,(x_j^* - E[x_j])$.

Our improvements of Kernel SHAP

We are concerned with the second part of Kernel SHAP, and will try to estimate $p(\boldsymbol{x}_{\bar{\mathcal{S}}} | \boldsymbol{x}_{\mathcal{S}} = \boldsymbol{x}^*_{\mathcal{S}})$ as good as possible, accounting for dependence.
We propose four approaches for estimating $p(\boldsymbol{x}_{\bar{\mathcal{S}}} | \boldsymbol{x}_{\mathcal{S}} = \boldsymbol{x}^*_{\mathcal{S}})$;

① assuming a **Gaussian distribution** for $p(\boldsymbol{x})$,

② assuming a **Gaussian copula distribution** for $p(\boldsymbol{x})$,

③ approximating $p(\boldsymbol{x}_{\bar{\mathcal{S}}} | \boldsymbol{x}_{\mathcal{S}} = \boldsymbol{x}^*_{\mathcal{S}})$ by an **empirical (conditional) distribution**,

④ a **combination of the empirical approach and either the Gaussian or the Gaussian copula** approach.

The empirical conditional approach is motivated by the idea that samples $(\boldsymbol{x}_{\bar{\mathcal{S}}}, \boldsymbol{x}_{\mathcal{S}})$ with $\boldsymbol{x}_{\mathcal{S}}$ close to $\boldsymbol{x}^*_{\mathcal{S}}$ are informative about the conditional distribution $p(\boldsymbol{x}_{\bar{\mathcal{S}}} | \boldsymbol{x}^*_{\mathcal{S}})$.

Experiments to evaluate our methods

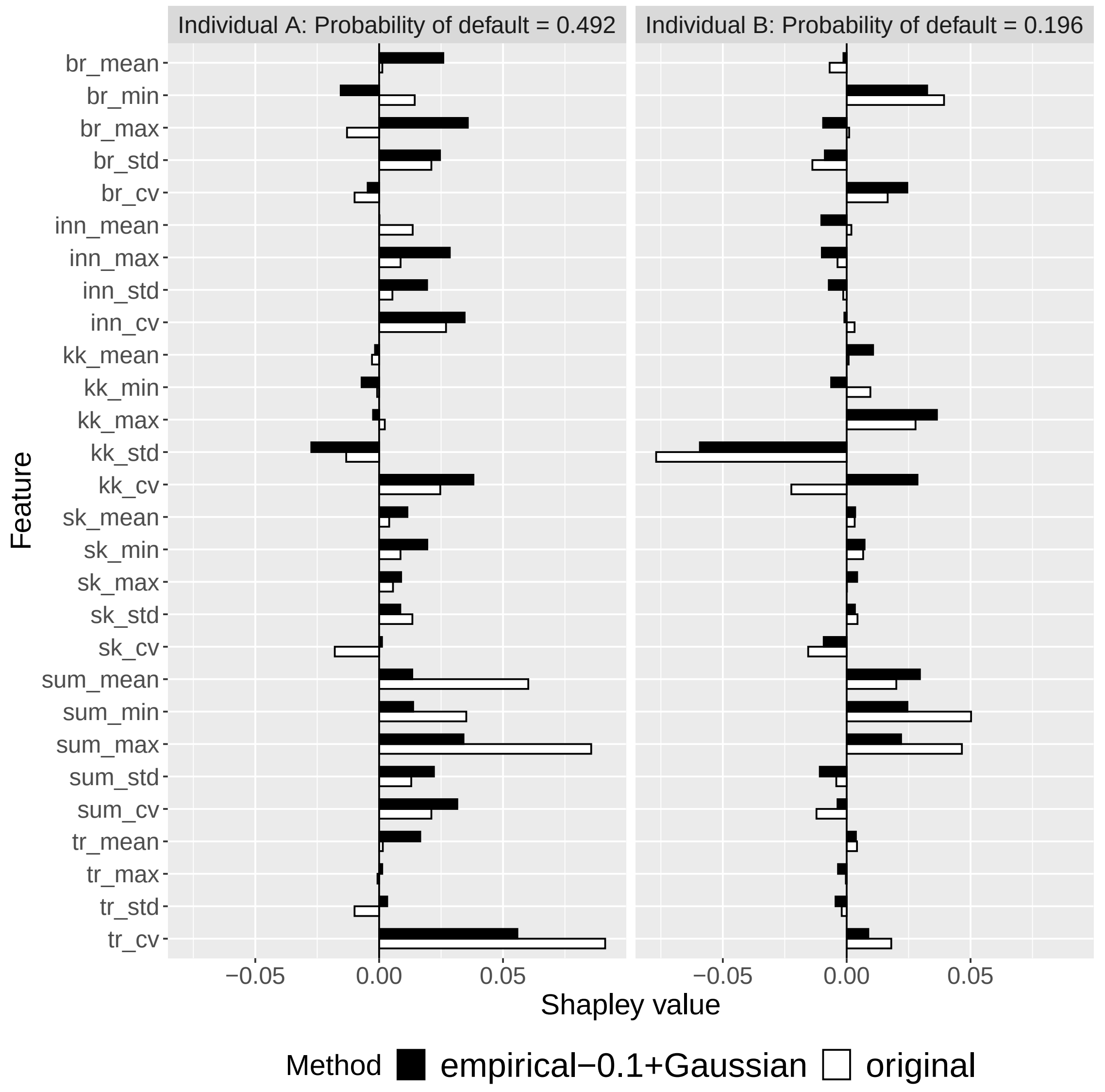
A problem with evaluating prediction explanation methods is that there generally is no ground truth. Hence, we turned to simulated data for which we may compute the true Shapley values. For the non-linear models, our methods clearly outperformed the TreeSHAP method [4], which tries to handle dependence between features.

The empirical conditional approach was superior when conditioning on a small number of the features. It was outperformed by the Gaussian and copula methods when conditioning on more features.

An example from finance

The data set consists of 28 features extracted from 6 transaction time series. It has previously been used for predicting mortgage default, and in the subsequent illustration, we have used Shapley values to explain two individuals' probabilities of default.

Explanations for two individuals



Shapley values for two persons in the real data set computed using our method and the original Kernel SHAP method. The Shapley values are quite different, illustrating that in the presence of feaure dependence, assuming independence can be misleading.

References

[1] Erik Štrumbel and Igor Kononenko.
Explaining prediction models and individual predictions with feature contributions.
Knowledge and Information Systems, 41:647–665, 2014.

[2] Scott M. Lundberg and Su-In Lee.
A Unified Approach to Interpreting Model Predictions.
In *Advances in Neural Information Processing Systems*, pages 4768–4777. Curram Associates Inc., 2017.

[3] Nikolai Sellereite and Martin Jullum.
shapr: An R-package for explaining machine learning models with dependence-aware Shapley values.
Journal of Open Source Software, 5(46):2027, 2020.

[4] Scott M. Lundberg and Su-In Lee.
Consistent feature attribution for tree ensembles.
In *Proceedings of the 34 th International Conference on Machine Learning*, pages 15–21. JMLR: W&CP, 2017.

BigInsight   NR Norsk Regnesentral NORWEGIAN COMPUTING CENTER