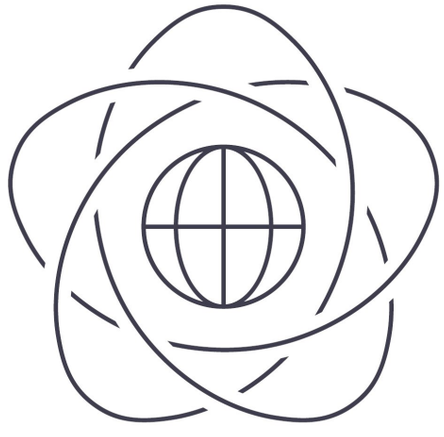


Case Studies 2022L

LIME Method and HW-I

Mar 24, 2022

Break-Down plot

Assume that prediction f(x) is an approximation of the expected value of the dependent variable Y given values of explanatory variables x.

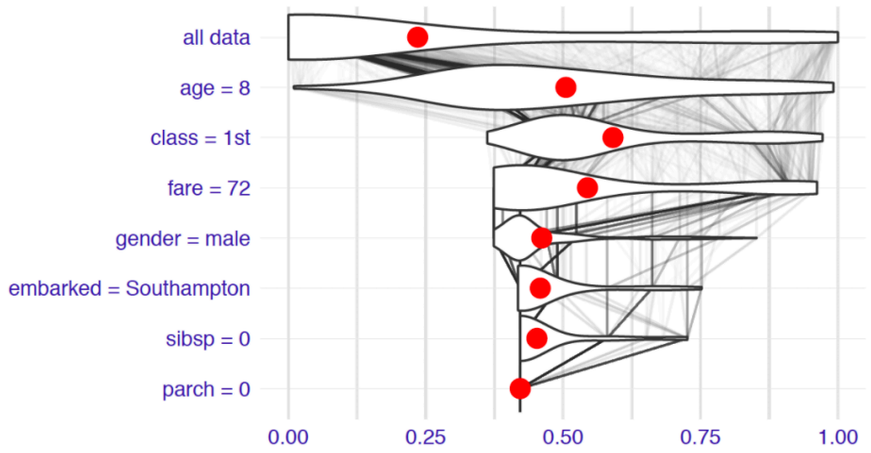
The underlying idea of BD plots is capture the contribution of an explanatory variable to the model's prediction by computing the shift in the expected value of Y, while fixing the values of other variables.

Let's look closer the following observation (Johnny D.) from Titanic dataset:

age: 8
class: 1st
fare: 72
gender: male
embarked: Southampton
sibsp (number of siblings/spouses aboard): 0
parch (number of parents/children aboard): 0

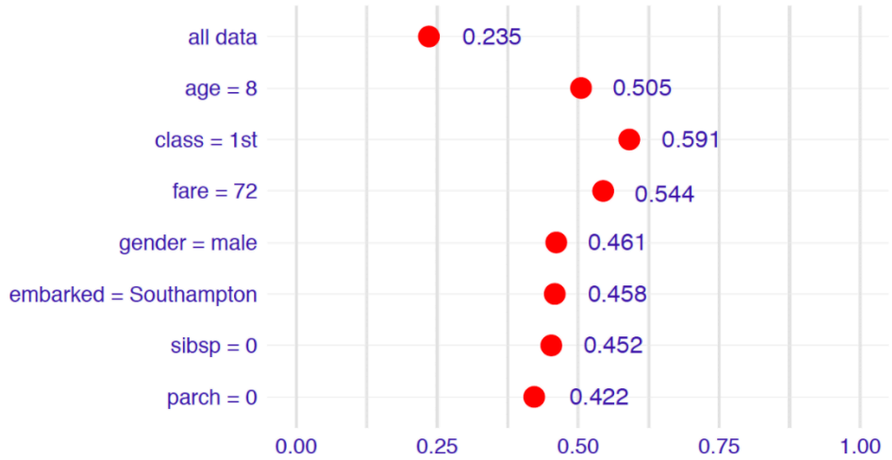
Break-Down plot

A)



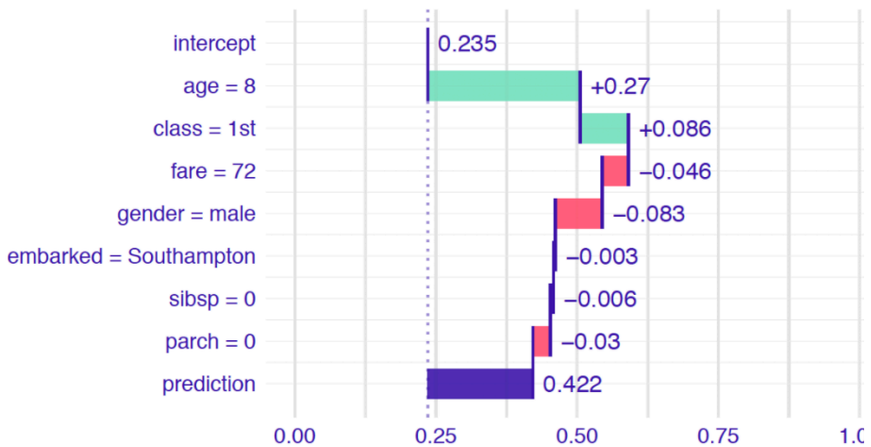
Violin plots of the predictions obtained from the model

B)



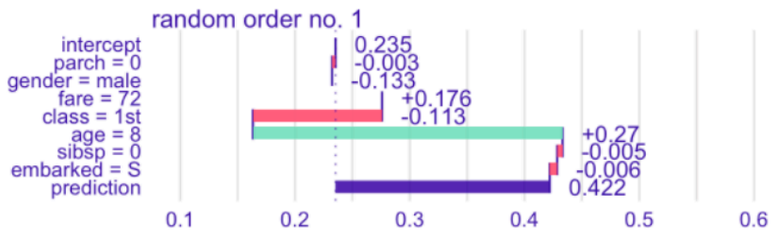
Mean values as an estimate of the expected value of the model's predictions

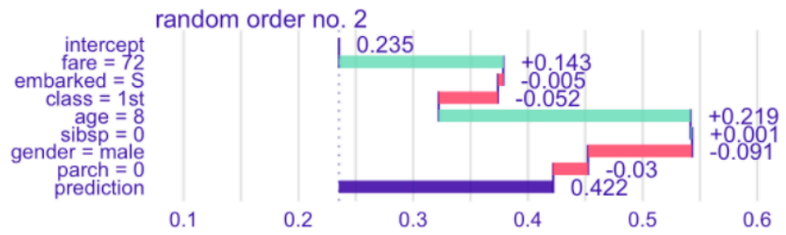
C)

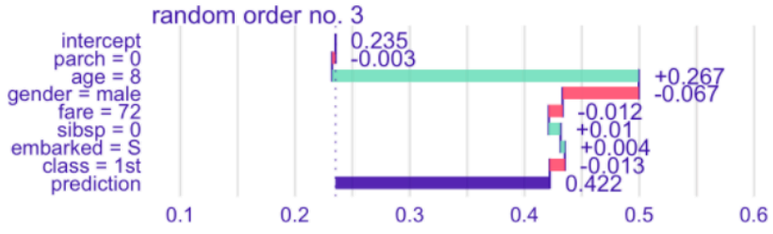


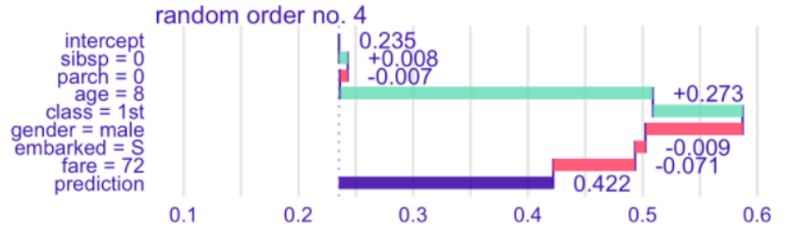
The contributions of the individual explanatory variables change the mean model's prediction

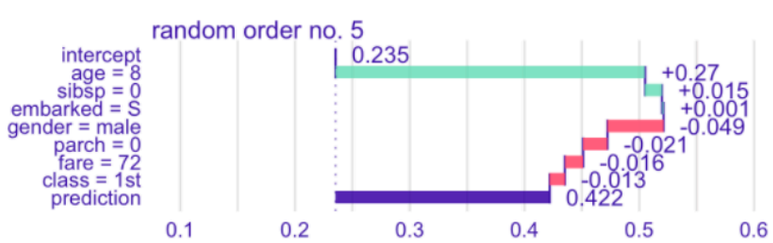
Break-Down plot

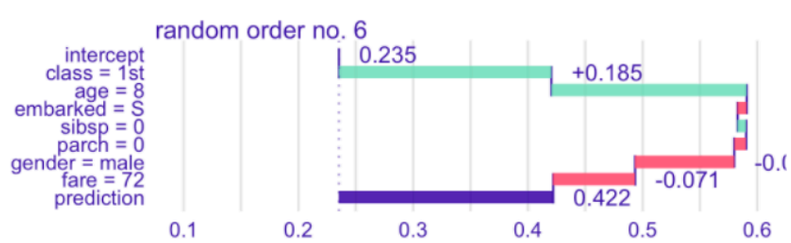


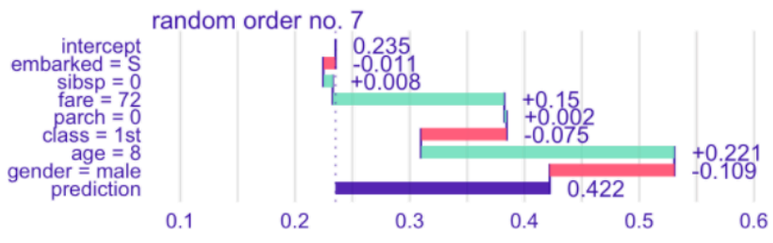


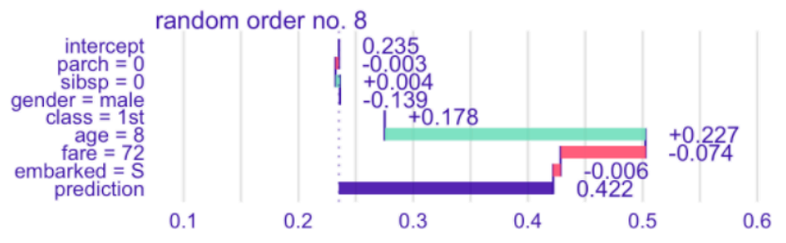


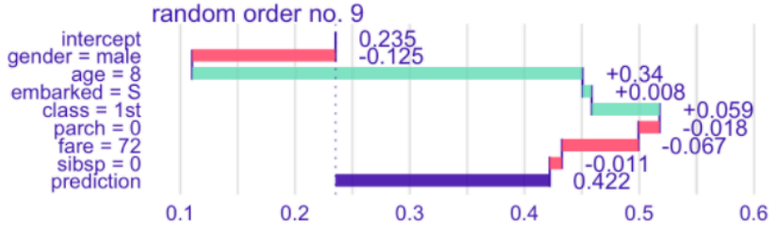


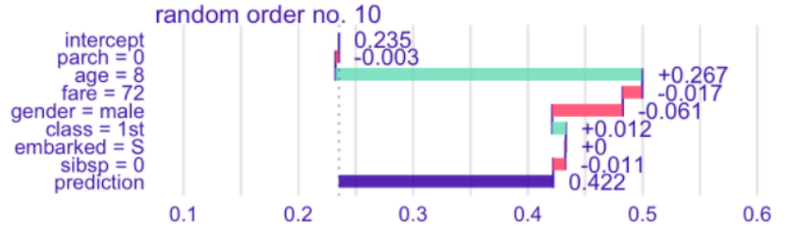












fare and class?

Pros and Cons

+ Model-agnostic

+ Easy-to-understand

+ Visualize

- Misleading for models including interactions*

- Complex for many explanatory variables

Shapley values

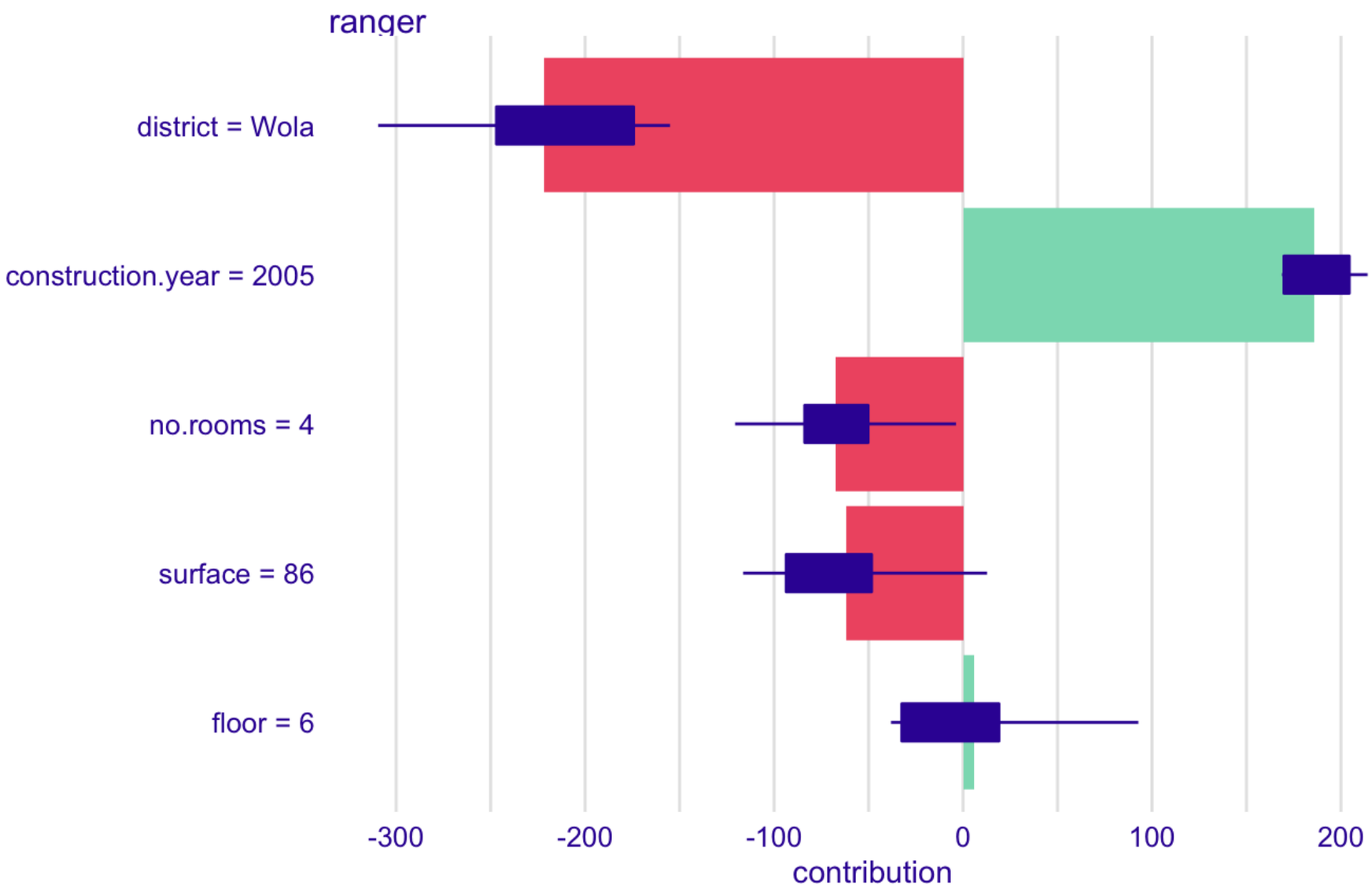
- A coalition of players cooperates and obtains a certain overall gain from the cooperation.

- Players are not identical, and different players may have different importance.

- Cooperation is beneficial, because it may bring more benefit than individual actions.

- The problem to solve is how to distribute the generated surplus among the players. Shapley values offer one possible fair answer to this question (Shapley 1953).

Shapley values



floor and surface?

Pros and Cons

+ Insensitive to influence of the ordering of the variables.

+ Solving the issues faced on using the BD plot.

- If the model is not additive, then the Shapley values may be misleading.

- The calculation of Shapley values is time-consuming for large models.

Local Interpretable Model-agnostic Explanations (LIME)

Break-down plots and Shapley values are most suitable for models with a small or moderate number of explanatory variables because they usually determine non-zero attributions for all variables in the model.

However, in domains like, for instance, genomics or image recognition, models with hundreds of thousands, or even millions, of explanatory variables are not uncommon.

The most popular example of such sparse explainers is the Local Interpretable Model-agnostic Explanations (LIME) method (Ribeiro et al., 2016).

Local Interpretable Model-agnostic Explanations (LIME)

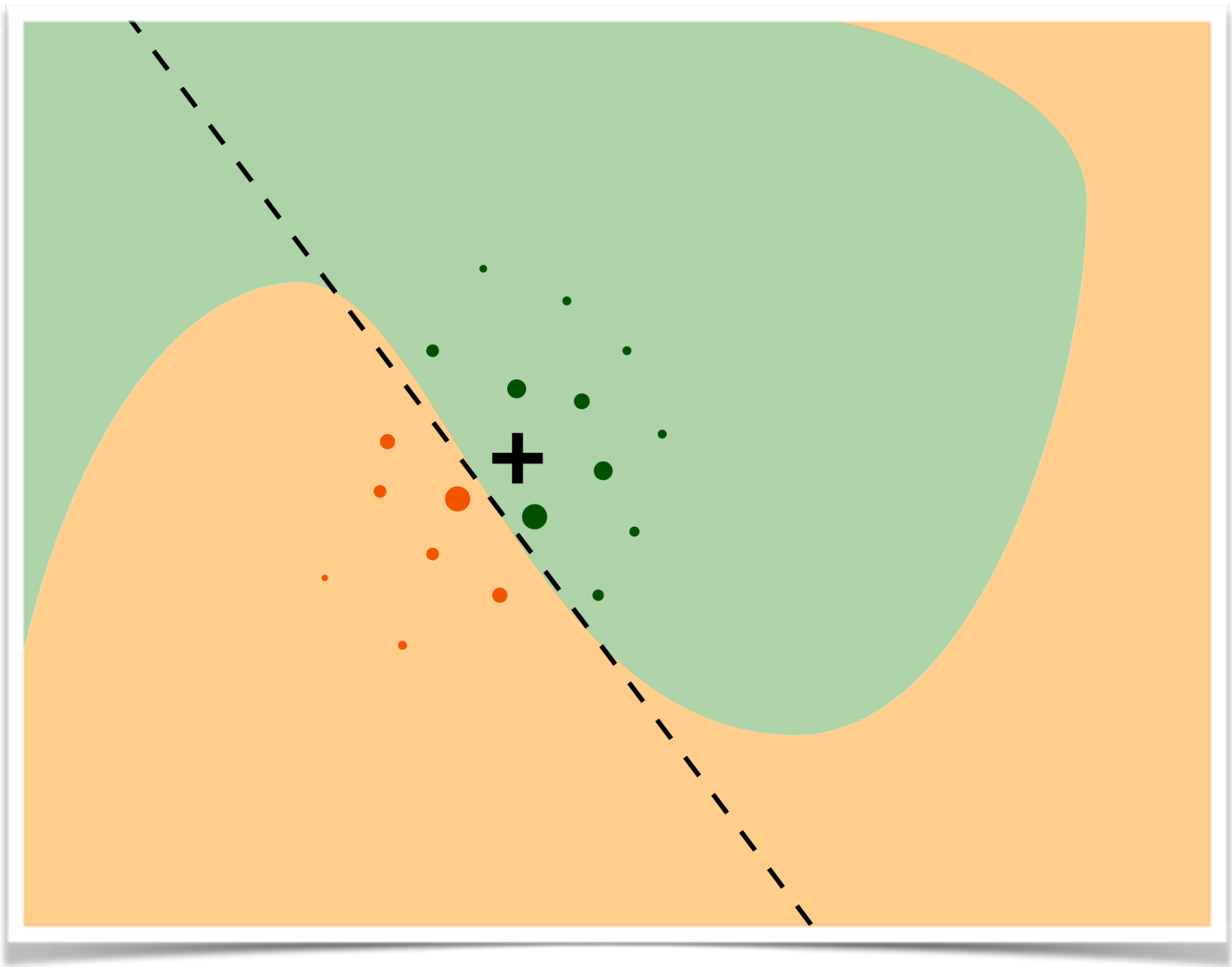


Figure 1: The idea behind the LIME approximation with a local glass-box model. The coloured areas correspond to decision regions for a complex binary classification model. The black cross indicates the instance (observation) of interest. Dots correspond to artificial data around the instance of interest. The dashed line represents a simple linear model fitted to the artificial data. The simple model “explains” local behavior of the black-box model around the instance of interest (Biecek and Burzykowski, 2020).

Local Interpretable Model-agnostic Explanations (LIME)

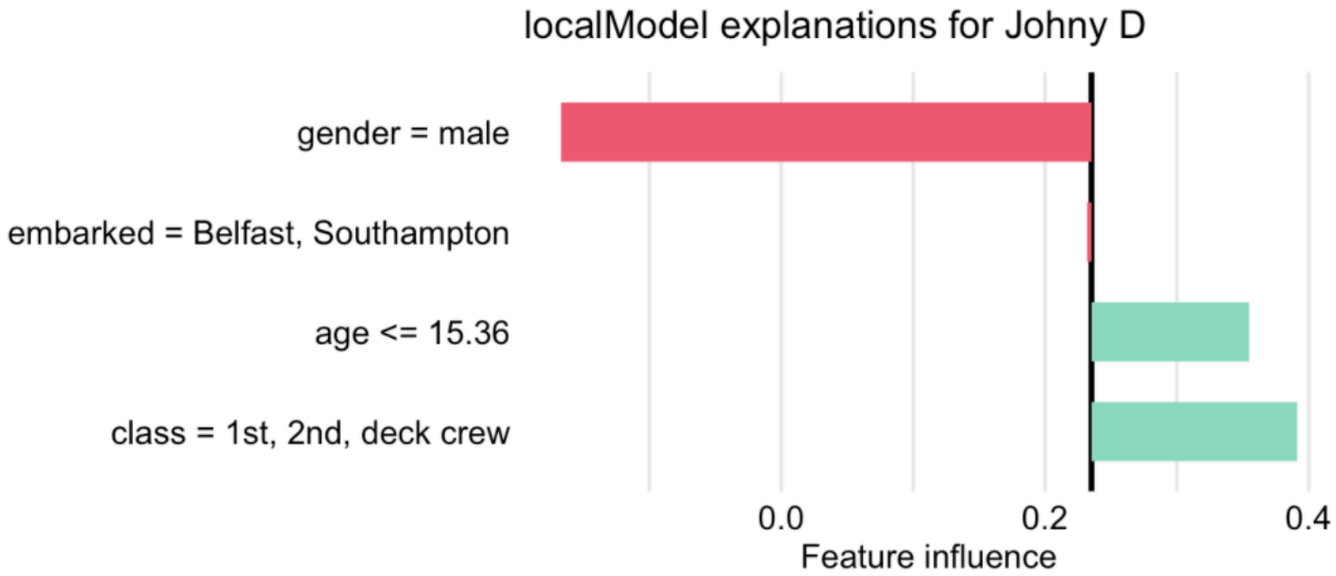
In LIME, we want to find a model that locally approximates a black-box model f() around the instance of interest x. Consider G of simple, interpretable models, like linear models. To find the required approximation, we minimize the loss function:

$$\hat{g} = \arg\min_{g \in \mathcal{G}} L\{f, g, \nu(\underline{x}_*)\} + \Omega(g),$$

The following algorithm is used to find an interpretable glass-box model g() that includes k most important, **interpretable**, explanatory variables:

```
Input: x* - observation to be explained
Input: N  - sample size for the glass-box model
Input: K  - complexity, the number of variables for the glass-box model
Input: similarity - a distance function in the original data space
1. Let x' = h(x*) be a version of x* in the lower-dimensional space
2. for i in 1...N {
3.   z'[i] <- sample_around(x')
4.   y'[i] <- f(z'[i])       # prediction for new observation z'[i]
5.   w'[i] <- similarity(x', z'[i])
6. }
7. return K-LASSO(y', x', w')
```

Local Interpretable Model-agnostic Explanations (LIME)



Pros and Cons

+ it is model-agnostic

+ offers an interpretable representation

+ provides local fidelity, i.e., the explanations are locally well-fitted to the black-box model.

- there have been various proposals for finding interpretable representations for continuous and categorical explanatory variables in case of tabular data. The issue has not been solved yet. This leads to different implementations of LIME, which use different variable transformation methods and, consequently, that can lead to different results.

- because the glass-box model is selected to approximate the black-box model, and not the data themselves, the method does not control the quality of the local fit of the glass-box model to the data. Thus, the latter model may be misleading.

- In high-dimensional data, data points are sparse. Defining a “local neighborhood” of the instance of interest may not be straightforward. Importance of the selection of the neighborhood is discussed, for example, by Alvarez-Melis and Jaakkola (2018). Sometimes even slight changes in the neighborhood strongly affect the obtained explanations.

References

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. ""Why should I trust you?": Explaining the Predictions of Any Classifier." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Kdd San Francisco, ca*, 1135–44. New York, NY: Association for Computing Machinery.

Alvarez-Melis, David, and Tommi S. Jaakkola. 2018. "On the Robustness of Interpretability Methods." *ICML Workshop on Human Interpretability in Machine Learning (WHI 2018)*, June. http://arxiv.org/abs/1806.08049.

Biecek, P., and Burzykowski, T. (2021) Explanatory Model Analysis, Chapman and Hall/CRC, New York.

Please feel free to send e-mail about your questions!

📩 mustafa.cavus@pw.edu.pl