The Journal of Open Source Software

# iml: An R package for Interpretable Machine Learning

**Christoph Molnar**[1], **Giuseppe Casalicchio**[1], and **Bernd Bischl**[1]

**1** Department of Statistics, LMU Munich

**DOI:** 10.21105/joss.00786

**Software**

- Review
- Repository
- Archive

**Submitted:** 19 June 2018
**Published:** 27 June 2018

## Summary

Complex, non-parametric models, which are typically used in machine learning, have proven to be successful in many prediction tasks. But these models usually operate as black boxes: While they are good at predicting, they are often not interpretable. Many inherently interpretable models have been suggested, which come at the cost of losing predictive power. Another option is to apply interpretability methods to a black box model after model training. Given the velocity of research on new machine learning models, it is preferable to have model-agnostic tools which can be applied to a random forest as well as to a neural network. Tools for model-agnostic interpretability methods should improve the adoption of machine learning.

`iml` is an R package (R Core Team 2016) that offers a general toolbox for making machine learning models interpretable. It implements many model-agnostic methods which work for any type of machine learning model. The package covers following methods:

- Partial dependence plots (Friedman 2001): Visualizing the learned relationship between features and predictions.
- Individual conditional expectation (Goldstein et al. 2015): Visualizing the learned relationship between features and predictions for individual instances of the data.
- Feature importance (Fisher, Rudin, and Dominici 2018): Scoring features by contribution to predictive performance.
- Global surrogate tree: Approximating the black box model with an interpretable decision tree.
- Local surrogate models (Ribeiro, Singh, and Guestrin 2016): Explaining single predictions by approximating the black box model locally with an interpretable model.
- Shapley value (Strumbelj et al. 2014): Explaining single predictions by fairly distributing the predicted value among the features.
- Interaction effects (Friedman, Popescu, and others 2008): Measuring how strongly features interact with each other in the black box model.

`iml` was designed to provide a class-based and user-friendly way to make black box machine learning models interpretable. Internally, the implemented methods inherit from the same parent class and share a common framework for the computation. Many of the methods are already implemented in other packages (e.g. (Greenwell 2017), (Goldstein et al. 2015), (Pedersen and Benesty 2017)), but the `iml` package implements all of the methods in one place, uses the same syntax and offers consistent functionality and outputs. `iml` can be used with models from the R machine learning libraries `mlr` and `caret`, but the package is flexible enough to work with models from other packages as well. Similar projects are the R package `DALEX` (Biecek 2018) and the Python package `Skater` (Choudhary, Kramer, and team 2018). The difference to `iml` is that the other two projects do not implement the methods themselves, but depend on other packages. `DALEX` focuses more on model comparison, and `Skater` additionally includes interpretable models and has less model-agnostic interpretability methods compared to `iml`.

The unified interface provided by the iml package simplifies the analysis and interpretation of black box machine learning learning models.

## Acknowledgements

This work is funded by the Bavarian State Ministry of Science and the Arts in the framework of the Centre Digitisation.Bavaria (ZD.B)

## References

Biecek, Przemyslaw. 2018. *DALEX: Descriptive mAchine Learning Explanations*. https://CRAN.R-project.org/package=DALEX.

Choudhary, Pramit, Aaron Kramer, and contributors datascience.com team. 2018. "Skater: Model Interpretation Library." https://doi.org/10.5281/zenodo.1198885.

Fisher, Aaron, Cynthia Rudin, and Francesca Dominici. 2018. "Model Class Reliance: Variable Importance Measures for any Machine Learning Model Class, from the "Rashomon" Perspective." http://arxiv.org/abs/1801.01489.

Friedman, Jerome H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics*. JSTOR, 1189–1232. https://doi.org/10.1214/aos/1013203451.

Friedman, Jerome H, Bogdan E Popescu, and others. 2008. "Predictive Learning via Rule Ensembles." *The Annals of Applied Statistics* 2 (3). Institute of Mathematical Statistics:916–54. https://doi.org/10.1214/07-AOAS148.

Goldstein, Alex, Adam Kapelner, Justin Bleich, and Emil Pitkin. 2015. "Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation." *Journal of Computational and Graphical Statistics* 24 (1):44–65. https://doi.org/10.1080/10618600.2014.907095.

Greenwell, Brandon M. 2017. "Pdp: An R Package for Constructing Partial Dependence Plots." *The R Journal* 9 (1):421–36. https://journal.r-project.org/archive/2017/RJ-2017-016/index.html.

Pedersen, Thomas Lin, and Michaël Benesty. 2017. *Lime: Local Interpretable Model-Agnostic Explanations*. https://CRAN.R-project.org/package=lime.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

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*, 1135–44. ACM. https://doi.org/10.1145/2939672.2939778.

Strumbelj, Erik, Igor Kononenko, Erik Štrumbelj, and Igor Kononenko. 2014. "Explaining prediction models and individual predictions with feature contributions." *Knowledge and Information Systems* 41 (3):647–65. https://doi.org/10.1007/s10115-013-0679-x.