

# dalex: Responsible Machine Learning with Interactive Explainability and Fairness in Python

Hubert Baniecki<sup>1</sup>

Wojciech Kretowicz<sup>1</sup>

Piotr Piatyszek<sup>1</sup>

Jakub Wisniewski<sup>1</sup>

Przemyslaw Biecek<sup>1,2</sup>

HUBERT.BANIECKI.STUD@PW.EDU.PL

WOJCIECH.KRETOWICZ.STUD@PW.EDU.PL

PIOTR.PIATYSZEK.STUD@PW.EDU.PL

JAKUB.WISNIEWSKI10.STUD@PW.EDU.PL

PRZEMYSLAW.BIECEK@PW.EDU.PL

<sup>1</sup>*Faculty of Mathematics and Information Science, Warsaw University of Technology*

*75 Koszykowa Street, Warsaw, Poland,*

<sup>2</sup>*Samsung Research & Development Institute, Poland*

## Abstract

The increasing amount of available data, computing power, and the constant pursuit for higher performance results in the growing complexity of predictive models. Their black-box nature leads to opaqueness debt phenomenon inflicting increased risks of discrimination, lack of reproducibility, and deflated performance due to data drift. To manage these risks, good MLOps practices ask for better validation of model performance and fairness, higher explainability, and continuous monitoring. The necessity of deeper model transparency appears not only from scientific and social domains, but also emerging laws and regulations on artificial intelligence. To facilitate the development of responsible machine learning models, we showcase **dalex**, a Python package which implements the model-agnostic *interface* for interactive model exploration. It adopts the design crafted through the development of various tools for responsible machine learning; thus, it aims at the *unification* of the existing solutions. This library's source code and documentation are available under open license at <https://python.drwhy.ai/>.

**Keywords:** machine learning, explainability, fairness, interactivity, responsible ml

## 1. Introduction

From the evolution of statistical modelling through data mining and machine learning to so-called artificial intelligence (AI), we arrived at the point where advanced systems support, or even surpass, humans in various predictive tasks. These algorithms are available for broad user-bases through numerous machine learning frameworks in Python like **scikit-learn** (Pedregosa et al., 2011), **tensorflow** (Abadi et al., 2016), **xgboost** (Chen and Guestrin, 2016) or **lightgbm** (Ke et al., 2017) to name just a few. Nowadays, we see an increase of concerns regarding explainability (Lipton, 2018; Miller, 2019) and fairness (Binns, 2018; Holstein et al., 2019) of machine learning predictive models in research and commercial domains. An increasing number of researchers discuss various needs and features for frameworks related to responsible machine learning (Barredo Arrieta et al., 2019; Gill et al., 2020). For us, the primary objective is combining three aspects of model exploration: explainability, fairness, and crucially for human-model dialogue (Abdul et al., 2018), interactivity.

Related software most notably include Python packages from these categories: **lime** (Ribeiro et al., 2016), **shap** (Lundberg and Lee, 2017), **pdpbox** (SauceCat, 2018), **interpret**

(Nori et al., 2019), `alibi` (Klaise et al., 2019), and `aix360` (Arya et al., 2020) which implement various explainability methods; `aif360` (Bellamy et al., 2018), `aequitas` (Saleiro et al., 2018), and `fairlearn` (Bird et al., 2020) which implement various fairness methods; responsible AI tools for `tensorflow` (Abadi et al., 2016), e.g. `witwidget` (Wexler et al., 2020), which produce interactive dashboards supporting machine learning (partially also addressed by `interpret` and `fairlearn`). All these leave a room for improvement in combining various methods, while also connecting them to ever-growing modelling and data frameworks through a uniform abstraction layer.

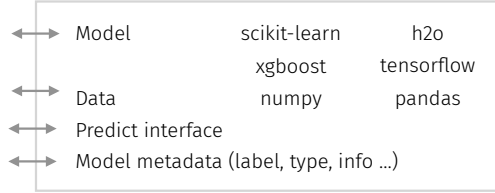
Unlike many of the proposed solutions, we strongly emphasise the construction of end-to-end software for facilitating the responsible approach to machine learning. The `dalex` package unifies various approaches and bridges the existing gap separating black-box models from explainability methods. Moreover, `dalex` brings numerous fairness metrics, and interactive model exploration dashboards, closer to the user. These combined motivate our contribution and article, where we preview our previous work in Section 2, introduce `dalex` in Section 3, and sketch the future work in Section 4.

## 2. Previous Work

This contribution builds upon the software for explainable machine learning presented by us in *"DALEX: Explainers for Complex Predictive Models in R"* (Biecek, 2018). Since `DALEX` version 0.2.5 there were two major releases, which expanded the toolkit of explainability methods, and performed a complete redesign of code, interface and charts for model visualizations. The number of users of this software grown significantly, which provided us with a great number of very valuable feature requests. Among others: (i) we created a taxonomy of model-agnostic explanations for machine learning predictive models (Biecek and Burzykowski, 2021); (ii) we prototyped `modelStudio` (Baniecki and Biecek, 2019), an extension of `DALEX`, which automatically produces a customizable dashboard, allowing for an interactive model exploration; (iii) we added support for multi-output predictive models and a growing number of machine learning frameworks in a language-agnostic manner. Further, we noticed that the visual model exploration goes beyond the area of explainability and also addresses such issues as fairness, contrastive comparisons and interactivity (Baniecki and Biecek, 2020). Based on these experiences, we implemented from scratch the `dalex` package for Python.

## 3. A Unified Interface for Responsible Machine Learning

The `dalex` Python package implements the main `dalex.Explainer` class to provide an abstract layer between distinct model API's (e.g. `scikit-learn` (Pedregosa et al., 2011), `tensorflow` (Abadi et al., 2016), `xgboost` (Chen and Guestrin, 2016), `h2o` (H2O.ai, 2020)) and data API's (e.g. `numpy` (Harris et al., 2020), `pandas` (Wes McKinney, 2010)), and the explainability and fairness methods. In Figure 1, we present the architecture of a unified interface for model-agnostic responsible machine learning with interactive explainability and fairness. These methods are divided into model-level techniques operating on the whole data (or its subset) and predict-level techniques operating on distinct observations from the data (or their neighbourhoods). Bounding of these methods to the one

**A. Explainer: Uniform abstraction over predictive models****B. Consistent grammar for model exploration**

```
import dalex as dx
explainer = dx.Explainer(model, X, y)

explanation = explainer.model_parts()
explanation.result
explanation.plot()

explainer.predict_parts(new_observation).result
explainer.predict_parts(new_observation).plot()
```

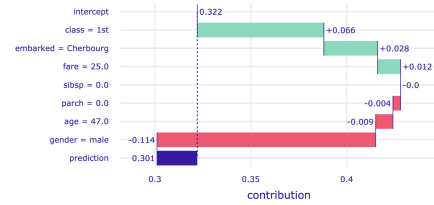
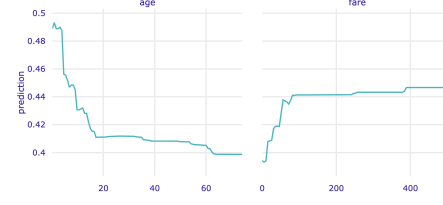
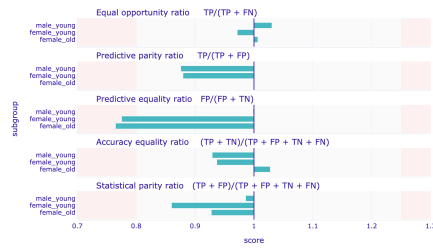
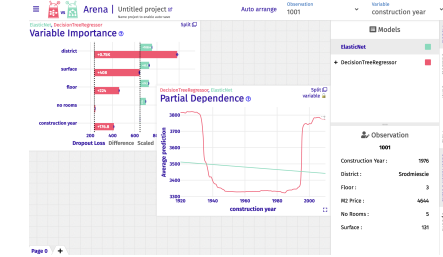
**C. Predict-level explanations****D. Model-level explanations****E. Fairness checks****F. Arena: Interactive cross-model exploration**

Figure 1: The **dalex** package is based on six pillars that support responsible machine learning modelling: A) The main **Explainer** class object, which serves as a uniform abstraction over predictive models and data API's in Python, B) A consistent set of methods for model exploration with explanation objects which calculate results, and plot them in a consistent way, C) Predict-level (local) explainability methods, D) Model-level (global) explainability methods, E) Fairness oriented methods, F) Interactive dashboard for contrastive model comparisons.

**dalex.Explainer** class gives a favourable user experience, where one can conveniently compute and return various explanation objects. All of them share the main **result** attribute, which is a **pandas.DataFrame**, and the **plot** method, which produces visualizations with the **plotly** package (Plotly Technologies Inc., 2015). The last takes multiple explanation objects allowing for an easy model comparison.

### 3.1 Model-level and predict-level explanations

Model-level and predict-level methods referenced in Figure 1 may return different objects depending on the **type** parameter: **predict** and **model\_performance** allow for an easy interference with the model basics, **predict\_parts** implements **iBreakDown** local variable

attributions and shapley values estimation, `model_parts` implements permutational variable importance, `predict_profile` implements Ceteris Paribus profiles, `model_profile` implements PDP, ALE and ICE profiles, `model_diagnostics` implements overall diagnostics of models' residuals, `model_surrogate` implements surrogate decision tree models which are effective to plot. Additionally, the `dalex.Explainer` abstract layer allows for the integration of other explanations, e.g. the `shap` (Lundberg and Lee, 2017) explanations into `predict_parts` and `model_parts` methods, and `lime` (Ribeiro et al., 2016) into `predict_surrogate`. All of these explanations are described in detail in *EMA* book (Biecek and Burzykowski, 2021) with `dalex` Python code examples.

### 3.2 Fairness check

The principles of responsible machine learning involve providing proper model accountability and bias detection (Barredo Arrieta et al., 2019; Gill et al., 2020). Because of regulations and guidelines, we see an increasing demand for easily accessible methods for checking model fairness (Binns, 2018; Holstein et al., 2019). This has resulted in the `fairness_check` method, which compares the most common fairness measures (Feldman et al., 2015; Verma and Rubin, 2018) and provides a detailed textual description of the group fairness analysis. It operates on fairness objects available through the `dalex.Explainer.model_fairness` method. In the same way as explanation objects, it contains the `result` attribute and `plot` method, which provides various visualizations depending on the `type` parameter. The individual fairness field is not as well established as the group one and there is still some discussion, and research to be done in case of individual fairness metrics.

### 3.3 Interactive and contrastive model exploration

The user-centred design of explainable (responsible) AI tools brings other emerging challenges discussed on the junction of AI and HCI domains (Abdul et al., 2018). The class `dalex.Arena` creates a live Arena dashboard (Piatyszek and Biecek, 2020) for model comparisons with all features available in the `dalex` package, including model explainability and fairness, and techniques for data exploration.

## 4. Conclusions and Future Work

In this article, we presented `dalex`, which builds upon and extends the DALEX R package to bring a unified interface for responsible machine learning into Python. This package is continuously developed, while the current stable version is v1.0 for Python 3.8 available at <https://python.drwhy.ai/>. Due to the comprehensive design of a uniform abstraction layer, our package allows for the addition of new machine learning frameworks into the responsible realm, which is not the case for most of the existing solutions. Additionally, with a clear-cut taxonomy of methods, there is a possibility to add new returning objects, which was well-proven within our previous contributions. Finally, the responsible machine learning domain aims to meet more principles than explainability and fairness (Barredo Arrieta et al., 2019); thus, next steps shall address security, safety, accountability, and privacy of machine learning models.

## Acknowledgments

We want to thank the <https://drwhy.ai/> community, users, researchers, and developers, for the continuous valid feedback over the past years. This work was financially supported by the (Poland) NCN Opus grant 2017/27/B/ST6/0130.

## References

- M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: A System for Large-Scale Machine Learning. In *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, OSDI’16, page 265–283, 2016.
- A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, and M. Kankanhalli. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI ’18, page 1–18, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356206. doi: 10.1145/3173574.3174156. URL <https://doi.org/10.1145/3173574.3174156>.
- V. Arya, R. K. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilovic, et al. AI Explainability 360: An Extensible Toolkit for Understanding Data and Machine Learning Models. *Journal of Machine Learning Research*, 21(130):1–6, 2020. URL <https://www.jmlr.org/papers/v21/19-1035.html>.
- H. Baniecki and P. Biecek. modelStudio: Interactive studio with explanations for ML predictive models. *Journal of Open Source Software*, 4(43):1798, Nov 2019. URL <https://doi.org/10.21105/joss.01798>.
- H. Baniecki and P. Biecek. The grammar of interactive explanatory model analysis. *arXiv (2005.00497)*, 2020. URL <https://arxiv.org/abs/2005.00497>.
- A. Barredo Arrieta, N. Diaz Rodriguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado González, S. Garcia, S. Gil-Lopez, D. Molina, V. R. Benjamins, R. Chatila, and F. Herrera. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion*, 2019. URL <https://doi.org/10.1016/j.inffus.2019.12.012>.
- R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, 2018. URL <https://arxiv.org/abs/1810.01943>.
- P. Biecek. DALEX: Explainers for Complex Predictive Models in R. *Journal of Machine Learning Research*, 19(84):1–5, 2018. URL <http://jmlr.org/papers/v19/18-416.html>.

- P. Biecek and T. Burzykowski. *Explanatory Model Analysis*. Chapman and Hall/CRC, New York, 2021. ISBN 9780367135591. URL <https://pbiecek.github.io/ema/>.
- R. Binns. Fairness in machine learning: Lessons from political philosophy. In *Conference on Fairness, Accountability and Transparency*, pages 149–159. PMLR, 2018.
- S. Bird, M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, and K. Walker. Fairlearn: A toolkit for assessing and improving fairness in AI. Technical Report MSR-TR-2020-32, Microsoft, May 2020. URL <https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/>.
- T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, 2016. URL <http://doi.acm.org/10.1145/2939672.2939785>.
- M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 259–268, 2015. URL <https://doi.org/10.1145/2783258.2783311>.
- N. Gill, P. Hall, K. Montgomery, and N. Schmidt. A Responsible Machine Learning Workflow with Focus on Interpretable Models, Post-hoc Explanation, and Discrimination Testing. *Information*, 11(3):137, 2020. URL <https://www.mdpi.com/2078-2489/11/3/137>.
- H2O.ai. *Python Interface for H2O*, 12 2020. URL <https://github.com/h2oai/h2o-3>. 3.32.0.2.
- C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del R’io, M. Wiebe, P. Peterson, P. G’erard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020. URL <https://doi.org/10.1038/s41586-020-2649-2>.
- K. Holstein, J. Wortman Vaughan, H. Daumé, M. Dudik, and H. Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI ’19, page 1–16, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300830. URL <https://doi.org/10.1145/3290605.3300830>.
- G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 3146–3154. Curran Associates, Inc., 2017. URL <https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf>.

- J. Klaise, A. Van Looveren, G. Vacanti, and A. Coca. *Alibi: Algorithms for monitoring and explaining machine learning models*, 2019. URL <https://github.com/SeldonIO/alibi>.
- Z. C. Lipton. The Mythos of Model Interpretability. *Queue*, 16(3):31–57, June 2018. URL <https://dl.acm.org/doi/abs/10.1145/3236386.3241340>.
- S. M. Lundberg and S.-I. Lee. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, pages 4765–4774. 2017. URL <http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf>.
- T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019. URL <https://doi.org/10.1016/j.artint.2018.07.007>.
- H. Nori, S. Jenkins, P. Koch, and R. Caruana. InterpretML: A Unified Framework for Machine Learning Interpretability. *arXiv*, 2019. URL <https://arxiv.org/pdf/1909.09223.pdf>.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- P. Piatyszek and P. Biecek. *Arena: universal dashboard for models exploration*, 12 2020. URL <https://arena.drwhy.ai/>. 0.3.0.
- Plotly Technologies Inc. Collaborative data science, 2015. URL <https://plot.ly>.
- M. T. Ribeiro, S. Singh, and C. Guestrin. “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*, page 1135–1144, 2016. URL <https://arxiv.org/pdf/1602.04938.pdf>.
- P. Saleiro, B. Kuester, A. Stevens, A. Anisfeld, L. Hinkson, J. London, and R. Ghani. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*, 2018.
- SauceCat. *PDPbox: python partial dependence plot toolbox*, 2018. URL <https://github.com/SauceCat/PDPbox>.
- S. Verma and J. Rubin. Fairness definitions explained. FairWare ’18, page 1–7, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450357463. URL <https://doi.org/10.1145/3194770.3194776>.
- Wes McKinney. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, pages 56 – 61, 2010. URL <https://doi.org/10.25080/Majora-92bf1922-00a>.
- J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson. The What-If Tool: Interactive Probing of Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):56–65, 2020. URL <https://doi.org/10.1109/TVCG.2019.2934619>.