Journal of Machine Learning Research 22 (2021) 1-7 Submitted 12/20; Revised 7/21; Published 9/21

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

Hubert Baniecki[1] HUBERT.BANIECKI.STUD@PW.EDU.PL
Wojciech Kretowicz[1] WOJCIECH.KRETOWICZ.STUD@PW.EDU.PL
Piotr Piatyszek[1] PIOTR.PIATYSZEK.STUD@PW.EDU.PL
Jakub Wisniewski[1] JAKUB.WISNIEWSKI10.STUD@PW.EDU.PL
Przemyslaw Biecek[1,2] PRZEMYSLAW.BIECEK@PW.EDU.PL
[1]*Faculty of Mathematics and Information Science, Warsaw University of Technology, Poland*
[2]*Samsung Research & Development Institute, Poland*

Editor: Joaquin Vanschoren

Abstract

In modern machine learning, we observe the phenomenon of *opaqueness debt*, which manifests itself by an increased risk of discrimination, lack of reproducibility, and deflated performance due to data drift. An increasing amount of available data and computing power results in the growing complexity of black-box predictive models. To manage these issues, good MLOps practice asks for better validation of model performance and fairness, higher explainability, and continuous monitoring. The necessity for deeper model transparency comes from both scientific and social domains and is also caused by emerging laws and regulations on artificial intelligence. To facilitate the responsible development of machine learning models, we introduce `dalex`, a Python package which implements a model-agnostic *interface* for interactive explainability and fairness. It adopts the design crafted through the development of various tools for explainable machine learning; thus, it aims at the *unification* of existing solutions. This library's source code and documentation are available under open license at `https://python.drwhy.ai`.

Keywords: explainability, fairness, interactivity, interpretability, responsible AI

1. Introduction

From the evolution of statistical modeling 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, there are increased concerns regarding the explainability (Lipton, 2018; Miller, 2019) and fairness (Binns, 2018; Holstein et al., 2019) of machine learning predictive models in research and commercial domains. A growing number of stakeholders 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 analysis: explainability, fairness, and crucially for human-model dialogue, interactivity (Abdul et al., 2018).

Related software most notably include Python packages from these three categories. `lime` (Ribeiro et al., 2016), `shap` (Lundberg and Lee, 2017), `pdpbox` (Jiangchun, 2018), `interpret` (Nori et al., 2019), `alibi` (Klaise et al., 2021), and `aix360` (Arya et al., 2020) implement various explainability methods; `aif360` (Bellamy et al., 2018), `aequitas` (Saleiro et al., 2018), and `fairlearn` (Bird et al., 2020) implement various fairness methods; moreover, responsible AI tools for `tensorflow` (Abadi et al., 2016), e.g. `witwidget` (Wexler et al., 2020), produce interactive dashboards supporting machine learning operations (this is also partially addressed by `interpret` and `fairlearn`). All these leave room for improvement in terms of the combining of various methods, while also connecting them to ever-growing modeling and data frameworks through a uniform abstraction layer.

Unlike many of the proposed solutions, we strongly emphasize the construction of end-to-end software for facilitating a responsible approach to machine learning. To achieve that, we focus on tabular data while there are frameworks specializing in other modalities, e.g. `innvestigate` (Alber et al., 2019). 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 analysis dashboards closer to the user. These factors motivate our article, in which 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 have been two major releases, which expanded the toolkit of explainability methods, and performed a complete redesign of code, interface and charts for model visualizations. Users provided us with a number of very valuable feature requests: (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 analysis (Baniecki and Biecek, 2020); (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 analysis goes beyond the area of explainability and also addresses such issues as fairness and interactive model comparisons. Based on these experiences, we implemented a Python package.

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 a whole dataset (or its subset) and predict-level techniques operating on distinct observa-

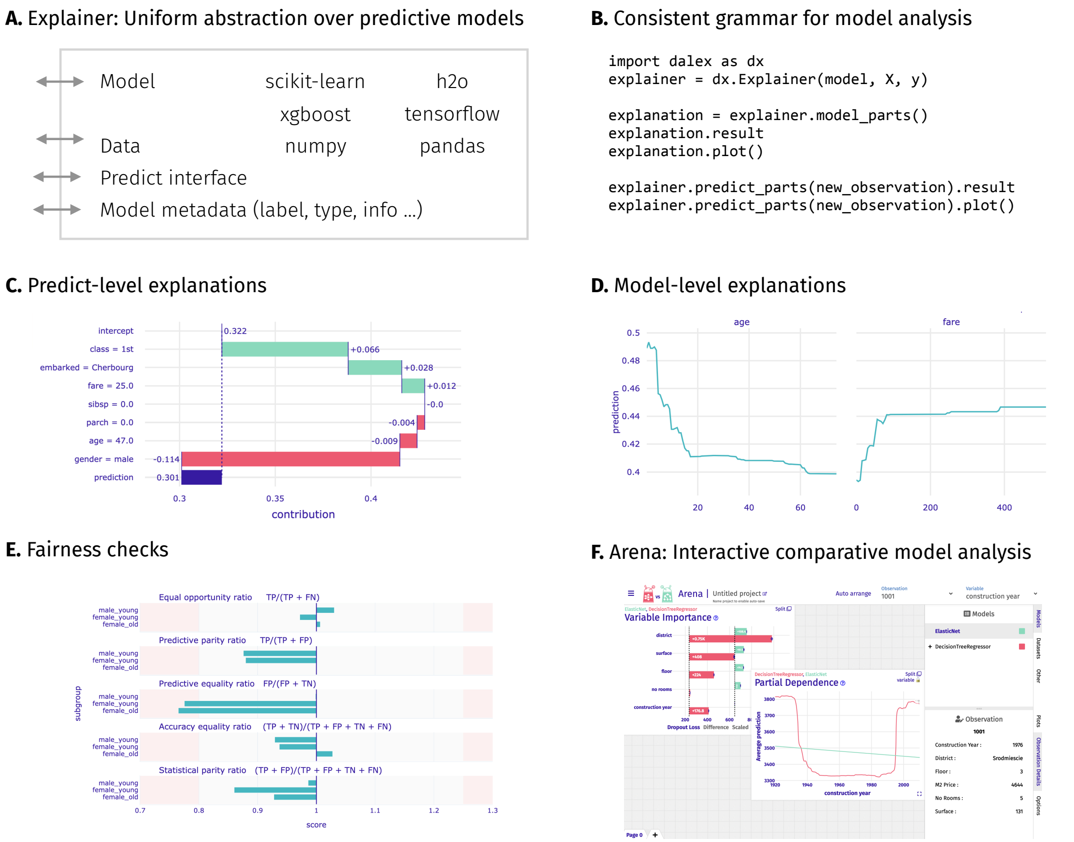

Figure 1: The dalex package is based on six pillars that support responsible machine learning modeling: **A.** The main Explainer class object, which serves as a uniform abstraction over predictive models and data API's in Python; **B.** A unified set of methods for model analysis with explanation objects that 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 comparative model analysis.

tions from data (or their neighbourhoods). The binding of these methods to the one `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 (Parmer and Kruchten, 2020). The latter takes multiple explanation objects, which allows for an easy model comparison.

Model-level and predict-level explanations. Explainability methods referenced in Figure 1 return different objects depending on the `type` parameter: `model_performance` and `predict` allow for 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 methods are described in detail in the *EMA* book (Biecek and Burzykowski, 2021) with `dalex` Python code examples.

Fairness checks. 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 can see an increasing demand for easily accessible methods to check model fairness (Binns, 2018; Holstein et al., 2019). Therefore, we implemented the `fairness_check` method, which compares the most common fairness measures based on the confusion matrix (Feldman et al., 2015; Verma and Rubin, 2018) and provides a detailed textual description of the group fairness analysis. It operates on a fairness object 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.

Interactive and comparative model analysis. 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; Miller, 2019). The `dalex.Arena` class creates an advanced live `Arena` dashboard (Piatyszek and Biecek, 2020) for model comparisons with all features available in the `dalex` package, including model explainability and fairness, moreover techniques for data exploration. These allow the juxtaposition of various visualizations for model and data analysis, which gives a complete view of the various models' behaviour. Notably, the dashboard can be saved into a local state to be loaded later — this overcomes the reproducibility crisis apparent in machine learning.

4. Conclusion and Future Work

In this article, we present `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 `1.3` for Python `3.9` is available at `https://python.drwhy.ai`. Due to the comprehensive design of a uniform abstraction layer, `dalex` allows for the convenient 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 the possibility to add new explanation objects and metrics, which was well-proven within our previous work. We further discuss such matters in the documentation and educational materials attached to this package.

We next aim to include into `dalex` explanations for groups of interacting variables, which is a highly influential concept in modern machine learning algorithms. There is research to be done towards adding a `predict_fairness` method, as the individual fairness field is not that well established. Overall, the responsible machine learning domain aims to address more principles than explainability and fairness (Barredo Arrieta et al., 2019); thus, the next steps shall address the accountability, robustness, and safety 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

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: A System for Large-Scale Machine Learning. *USENIX Conference on Operating Systems Design and Implementation*, pages 265–283, 2016.

Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. *CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2018.

Maximilian Alber, Sebastian Lapuschkin, Philipp Seegerer, Miriam Hägele, Kristof T. Schütt, Grégoire Montavon, Wojciech Samek, Klaus-Robert Müller, Sven Dähne, and Pieter-Jan Kindermans. iNNvestigate Neural Networks! *Journal of Machine Learning Research*, 20(93):1–8, 2019.

Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilovic, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John T. Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yunfeng Zhang. AI Explainability 360: An Extensible Toolkit for Understanding Data and Machine Learning Models. *Journal of Machine Learning Research*, 21(130):1–6, 2020.

Hubert Baniecki and Przemyslaw Biecek. modelStudio: Interactive studio with explanations for ML predictive models. *Journal of Open Source Software*, 4(43):1798, 2019.

Hubert Baniecki and Przemyslaw Biecek. The Grammar of Interactive Explanatory Model Analysis. *arXiv preprint arXiv:2005.00497*, 2020.

Alejandro Barredo Arrieta, Natalia Diaz Rodriguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado González, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, V. Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2019.

Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. AI

Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. *arXiv preprint arXiv:1810.01943*, 2018.

Przemyslaw Biecek. DALEX: Explainers for Complex Predictive Models in R. *Journal of Machine Learning Research*, 19(84):1–5, 2018.

Przemyslaw Biecek and Tomasz Burzykowski. *Explanatory Model Analysis*. Chapman and Hall/CRC, New York, 2021. ISBN 9780367135591. URL https://pbiecek.github.io/ema.

Reuben Binns. Fairness in Machine Learning: Lessons from Political Philosophy. *Conference on Fairness, Accountability and Transparency*, 81:149–159, 2018.

Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. Fairlearn: A toolkit for assessing and improving fairness in AI. Technical Report MSR-TR-2020-32, Microsoft, 2020.

Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and Removing Disparate Impact. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268, 2015.

Navdeep Gill, Patrick Hall, Kim Montgomery, and Nicholas Schmidt. A Responsible Machine Learning Workflow with Focus on Interpretable Models, Post-hoc Explanation, and Discrimination Testing. *Information*, 11(3):137, 2020.

H2O.ai. *Python Interface for H2O*, 2020. URL https://github.com/h2oai/h2o-3.

Charles R. Harris, K. Jarrod Millman, St'efan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fern'andez del R'ıo, Mark Wiebe, Pearu Peterson, Pierre G'erard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825): 357–362, 2020.

Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? *CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2019.

Li Jiangchun. *PDPbox: python partial dependence plot toolbox*, 2018. URL https://github.com/SauceCat/PDPbox.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*, 30:3146–3154, 2017.

Janis Klaise, Arnaud Van Looveren, Giovanni Vacanti, and Alexandru Coca. Alibi Explain: Algorithms for Explaining Machine Learning Models. *Journal of Machine Learning Research*, 22(181):1–7, 2021.

Zachary C. Lipton. The Mythos of Model Interpretability. *Queue*, 16(3):31–57, 2018.

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

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. InterpretML: A Unified Framework for Machine Learning Interpretability. *arXiv preprint arXiv:1909.09223*, 2019.

Chris Parmer and Nicolas Kruchten. *plotly: An open-source, interactive data visualization library for Python*, 2020. URL `https://github.com/plotly/plotly.py`.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

Piotr Piatyszek and Przemyslaw Biecek. *Arena: universal dashboard for model exploration*, 2020. URL `https://arena.drwhy.ai/`.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T. Rodolfa, and Rayid Ghani. Aequitas: A Bias and Fairness Audit Toolkit. *arXiv preprint arXiv:1811.05577*, 2018.

Sahil Verma and Julia Rubin. Fairness Definitions Explained. *International Workshop on Software Fairness*, pages 1–7, 2018.

Wes McKinney. Data Structures for Statistical Computing in Python. *Python in Science Conference*, pages 56–61, 2010.

James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viegas, and Jimbo Wilson. The What-If Tool: Interactive Probing of Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):56–65, 2020.