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

Hubert Baniecki1
Wojciech Kretowicz1
Piotr Piatyszek1
Jakub Wisniewski1
Przemyslaw Biecek1,2

1Faculty of Mathematics and Information Science, Warsaw University of Technology, Poland
2Samsung 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.

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


