Fairkit-learn: A Fairness Evaluation and Comparison Toolkit

Brittany Johnson
George Mason University
Fairfax, VA, USA
johnsonb@gmu.edu

Yuriy Brun
University of Massachusetts Amherst
Amherst, MA, USA
brun@cs.umass.edu

1 INTRODUCTION
Software engineering and data scientists, more and more, use data to train machine learning models as part of software systems. Not only is data-driven software becoming more pervasive, it is being adopted in contexts where unexpected outcomes can have detrimental impact.

While there is potential for data-driven software to improve our way of life, recent studies suggest that societal biases are in the data we use when training machine learning models, which leads to technological biases [12, 27]. YouTube makes more mistakes when rendering closed captions for female voices [22, 33]. E-commerce software has showcased bias in their services and discounts [16, 25]. Facial recognition software has difficulty accurately recognizing female and non-white faces [6, 17, 21]. Unfortunately, the list goes on and on.

One way to support data scientists in addressing bias in machine learning is by providing tools that can help them reason about the various considerations that come with training high quality, unbiased models. To this end, we developed fairkit-learn, an open source Python toolkit for evaluating and comparing machine learning models with respect to fairness and other quality metrics [20]. An evaluation of fairkit-learn found that it does in fact support the ability to find models that are both fair and high quality, and improves the ability to do so over scikit-learn and AI Fairness 360. This paper outlines the components of fairkit-learn and how it can be used to automatically (and with little overhead for the user) train, evaluate, and compare a large number of model configurations.

Next, Section 2 describes fairkit-learn and Section 3 details how it can be used to help develop fair software. Section 4 places our research in the context of related work, and Section 5 summarizes our contributions. A video of fairkit-learn in action is available at https://youtu.be/ZC_deJnI9xs/.

2 FAIRKIT-LEARN: MODEL EVALUATION AND COMPARISON
Figure 2 outlines the steps involved when using fairkit-learn. Along with the dataset being used for model training, the inputs to fairkit-learn are models, hyperparameters, metrics, protected attribute, classification threshold, and pre- and post-processing algorithms. In this section, we discuss the various components of fairkit-learn and how it uses these inputs.

Integrated machine learning tools. Fairkit-learn is built on top of scikit-learn and AIF360 [19, 30]. Given scikit-learn is a foundational machine learning toolkit, we wanted to make sure fairkit-learn could interface with its algorithms and metrics. We integrated AIF360, which also builds on top of scikit-learn, to provide fairkit-learn’s fairness-related functionality.
Fairkit-learn supports all of scikit-learn’s and AIF360’s algorithms and metrics, including all of AIF360’s bias mitigating algorithms. Fairkit-learn is currently capable of working with over 70 definitions of fairness, including:

- **Disparate impact**, which measures if a model treats similarly the same fraction of individuals of each group [10, 14, 37].
- **Demographic parity**, also called statistical parity and group fairness, which measures if a model’s predictions are statistically independent of the attribute with respect to which the model is fair [8, 11].
- **Equal opportunity difference**, which requires that false negative rates among groups are equal [10, 15].
- **Causal fairness**, also called counterfactual fairness, requires classifiers to predict the same outcome for two individuals that differ only in protected attributes, and are otherwise identical [12, 23].

Fairkit-learn also provides extension points for including additional metrics and algorithms. Fairkit-learn builds on the contributions of scikit-learn and AI Fairness 360 by providing the following unique features, which we discuss in more detail next:
• An automated model search capable of evaluating thousands of machine learning models with respect to two quality and/or fairness metrics simultaneously.
• An interactive visualization that allows users to explore and compare a small, Pareto-optimal set of models for each set of metrics selected.

**Model search.** Unlike existing tools, fairkit-learn allows users to find models that best balance fairness with other quality concerns across as many models as the user can computationally support (the more models you want to assess, the more memory and power needed). Figure 1 shows an example set of inputs used for fairkit-learn’s ModelSearch. The required inputs for fairkit-learn’s model search are:

- **Models.** Fairkit-learn requires the specification of at least one model (*models* in Figure 1) to perform the grid search, but users can specify as many models as their computational resources will allow.
- **Metrics.** Along with models, the user must specify the metrics fairkit-learn should use to evaluate each model configuration (*metrics* in Figure 1). All of fairkit-learn’s current metrics are stored in the `UnifiedMetricLibrary`.
- **Hyperparameters.** This is an optional input, where the user can specify value ranges for each hyperparameter of each model (*hyperparameters* in Figure 1). If no hyperparameter ranges are specified, fairkit-learn will use default model configurations in the grid search.
- **Thresholds.** Users need to specify the probabilistic threshold required for positive binary classification (*thresholds* in Figure 1). For example, a threshold of 0.8 will consider any prediction with \( \geq 0.7 \) probability to be favorable.
- **Pre- and post-processing algorithms.** Lastly, users can add any pre- and post-processing algorithms to include in the grid search (*preprocessors* and *postprocessors* in Figure 1).

Once the search is done, results are written to a .csv file that is used to render a visualization of the results from the grid search.

**Search visualization.** One of the ways that users can process the grid search results is presented through the interactive visualization provided by fairkit-learn, as shown in Figure 3. The plot that is shown was rendered using the results from the search in Figure 1. The fairkit-learn visualization allows for viewing of Pareto-optimal models for any two metrics by selecting those metrics from the checklist in the bottom left corner of Figure 3 and choosing those metrics as the X and Y axes in the dropboxes in the upper left corner. Users can also access all of the search results, including non-optimal models, by selecting all the metrics in the checklist. To toggle between models and metrics being displayed in the visualization, users can select the X and Y axes along with selecting (or de-selecting) models to include using the different color model buttons (e.g., the magenta AdversarialDebiasing button). Users can export the results of their search as a plot that comes with a JSON file that describes the plot.

### 3 USING FAIRKIT-LEARN TO FIND OPTIMAL MODELS

Fairkit-learn is an open source Python toolkit that supports interactive evaluation and comparison of machine learning models for fairness and other quality metrics simultaneously. It can evaluate thousands of models produced by multiple machine learning algorithms, hyperparameters, and data permutations, and compute then visualize a small Pareto-optimal set of models that provide an optimal balance between fairness and quality. Data scientists can then iterate, improving their models and evaluating them using fairkit-learn. Instructions for installing fairkit-learn, along with a tutorial implemented in a Jupyter notebook, can be found here: https://go.gmu.edu/fairkit-learn

To better understand how a data scientist could use fairkit-learn to train, evaluate, and compare machine learning models, let us look back at the search written in Figure 1. Here, the user wants to compare the resulting models from three algorithms: LogisticRegression, RandomForestClassifier, and AdversarialDebiasing. In this example, she is using the COMPAS recidivism dataset, which contains recidivism data for Browards County between the years 2013 and 2014 [28]. Let us imagine that the user wants to explore models that best balance accuracy and fairness; one could choose any metric for each of these concerns, however, she cares about the accuracy_score and the disparate_impact.

While the user has only entered three learning algorithms, fairkit-learn will train approximately 80 different models by using the hyperparameter value ranges specified to vary the hyperparameter values in the grid-search. The grid-search will produce a substantially smaller subset of models (in this case, the output includes only 7 models) that make up the Pareto-optimal set. This smaller set of models is then presented to the user in a visualization that plots the models with respect to the metrics chosen by the user; in this case, we have disparate impact (a fairness metric) on the x-axis and accuracy (a quality metric) on the y-axis.

The final set of optimal models shown in the visualization are models for which improving fairness would decrease accuracy and vice versa (the Pareto-optimal set). For the search shown in Figure 1, the interactive visualization in Figure 3 makes it easy to come realize that (1) given this dataset (COMPAS), model fairness and accuracy are often in opposition – in other domains they may be complementary, (2) we can achieve a large increase in fairness (69% compared to 45%) if we sacrifice some accuracy (63% compared to 68%), (3) while adversarial debiasing (magenta) is a fairness-aware algorithm, it produces less fair but slightly more accurate models in comparison to random forest classifier models (orange) that produce more fair models with a small decrease in accuracy, and (4) with respect to both accuracy and fairness, logistic regression models (purple) tend to be more balanced than the other two algorithms with accuracy that is comparable to adversarial debiasing but increased fairness.

Users can dig deeper into the search results by hovering over the data points in the plot to get more information on the Pareto-optimal models found (e.g., the hyperparameter values for that configuration), opening the .csv file used to render the visualization, which includes all models included in the search, or exporting the plot and using the accompanying JSON file to examine optimal models.
User Evaluation. To evaluate the potential for fairkit-learn to be useful in practice, we conducted a user study with 54 graduate and undergraduate students with varying experience training machine learning models [20]. We asked participants to complete a series of tasks involving training and evaluating machine learning models for fairness and accuracy. Along with using fairkit-learn, we asked participants to complete the same tasks using scikit-learn, the state-of-the-art in training machine learning models, and AIF360, one of them more recognized model fairness evaluation toolkits. We found that participants selected fairer, more accurate models when using fairkit-learn. When trying to balance fairness and accuracy, fairkit-learn was able to find models that are high performing and generally more fair than the models found by AIF360 and scikit-learn.

4 RELATED WORK

Typically, machine learning model performance is evaluated using accuracy metrics. scikit-learn [26], one of the most common tools used for training and evaluating machine learning models, provides engineers with a variety of machine learning algorithms and various metrics for evaluating models for performance. While scikit-learn is useful for training and evaluating models based on their performance, there is no built in functionality for measuring model fairness or mitigating bias. But fairness of machine learning models plays an important role in software that uses such models [5].

There do exist tools designed to help engineers reason about fairness in their machine learning models [1, 3, 36]. FairML supports the detection of unintended discrimination in predictive models by automatically determining the relative significance of model inputs to outcomes [1]. Another solution, Fairway, combines pre-processing and in-processing algorithms to remove bias from training data and models [9].

Fairkit-learn supports evaluating models with respect to two metrics simultaneously via an interactive visualization of Pareto-optimal models. Most related to our contribution is Microsoft’s Fairlearn, a Python toolkit that uses interactive visualizations to support evaluating model fairness and fairness-performance trade-offs [4]. While both fairkit-learn and Fairlearn provide similar functionality to accomplish similar goals, fairkit-learn has a couple of unique features. First, fairkit-learn was built using state-of-the-art machine learning and fairness libraries increasing ease of integration into existing workflows. Also, currently Fairlearn is only capable of measuring group discrimination while fairkit-learn supports any definition of fairness in AI Fairness 360 (or that the user would like to add or implement). Lastly, fairkit-learn only returns Pareto-optimal models.

Google developed the What-If Tool to support the analysis and understanding of machine learning models without having to write code [36]. When given a TensorFlow model and a dataset, the What-If Tool visualizes the dataset, allows for editing of individuals in the dataset, shows the effects of dataset modification, performs counterfactual analysis, and evaluates models based on performance and fairness. Farea, a model behavior mutation approach, is another intervention that supports measuring and evaluating fairness-accuracy trade-offs when using machine learning bias mitigation methods [18].

AI Fairness 360 (AIF360), a Python tool suite for evaluating model fairness and performance [3], includes fairness metrics, metric explanations, and bias mitigation algorithms for datasets and models. AIF360 is designed to be extensible and accessible to data scientists and practitioners. Also similar is FairVis, a visual analytics system that supports exploring fairness and performance with respect to certain subgroups in a dataset [7]. Similar to fairkit-learn, FairVis uses visualizations to support this exploration.

Some machine learning methods, known as Seldonian algorithms, provide high-confidence guarantees that learned models enforce user-specified fairness properties, even when applied to unseen data [24, 34]. These guarantees can even extend to settings when the distribution of the training data is different from that of the data to which the model is applied [13].

There also exist tools designed to support engineers ability to test their software for fairness [2, 12, 31, 35, 38, 39]. Themis, a software fairness testing tool, was the first of its kind [2, 12]. Themis automatically generates tests that help engineers detect and measure causal and group discrimination. Fairness testing can be made more efficient in finding inputs that exhibit bias [35, 38, 39], and can be driven by a grammar [31].

While there exists tools that can help engineers evaluate models for fairness and performance and measure software bias, fairkit-learn works with existing tools to help engineers find Pareto-optimal models that balance fairness and performance and provides an interactive visualization that makes it quicker and easier to explore the effects of different model configurations.

5 CONTRIBUTIONS

We presented fairkit-learn, a novel open-source toolkit designed to support the evaluation and comparison of machine learning models across multiple dimensions, such as fairness and performance. We described how to use fairkit-learn to train, evaluate, and compare machine learning models using its interactive visualization interface. We outline the potential for fairkit-learn to be beneficial in practice based on results from a controlled user study, demonstrating that fairkit-learn is an effective tool for helping data scientists understand the fairness-quality landscape.

ACKNOWLEDGMENTS

Jesse Bartola, Rico Angell, Sam Witty, Stephen J. Giguere, and Katherine A. Keith helped build and evaluate fairkit-learn. This work is supported by the National Science Foundation under grant no CCF-1763423, and by Google, Meta Platforms, and Kosai.ai.

REFERENCES


