Data vs. Model Machine Learning Fairness Testing

An Empirical Study

Shome, Arumoy; Cruz, Luís; Van Deursen, Arie

DOI
10.1145/3639478.3643121

Publication date
2024

Document Version
Final published version

Published in
Proceedings - 2024 ACM/IEEE 46th International Conference on Software Engineering

Data vs. Model Machine Learning Fairness Testing: An Empirical Study

Arumoy Shome
Delft University of Technology
Delft, Netherlands
a.shome@tudelft.nl

Luís Cruz
Delft University of Technology
Delft, Netherlands
l.cruz@tudelft.nl

Arie van Deursen
Delft University of Technology
Delft, Netherlands
arie.vandeursen@tudelft.nl

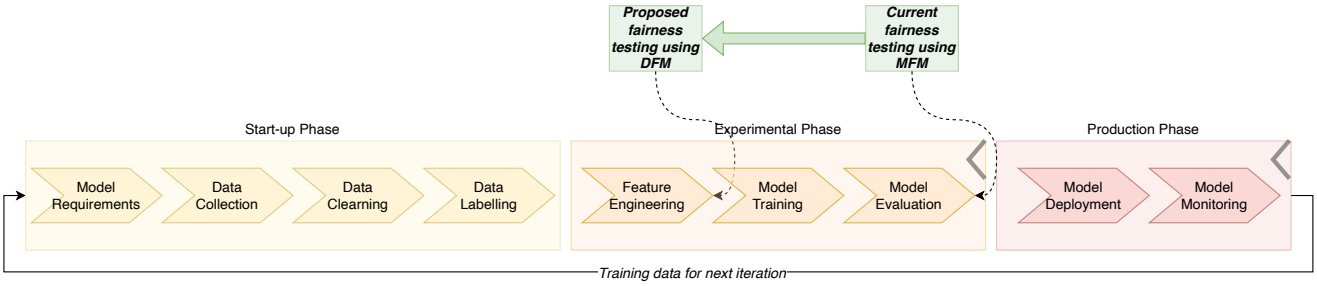

Figure 1: Stages of the ML Lifecycle (adopted from [1, 3]). Three distinct phases of the lifecycle are marked by different colours. Stages in the experimental and production phases may loop back to any prior stages, indicated by the large grey arrows. The location of fairness testing using DFM and MFM are marked by the green labels. The green arrow depicts the shift proposed by this study in ML fairness testing.

ABSTRACT

Although several fairness definitions and bias mitigation techniques exist in the literature, all existing solutions evaluate fairness of Machine Learning (ML) systems after the training stage. In this paper, we take the first steps towards evaluating a more holistic approach by testing for fairness both before and after model training. We evaluate the effectiveness of the proposed approach and position it within the ML development lifecycle, using an empirical analysis of the relationship between model dependent and independent fairness metrics. The study uses 2 fairness metrics, 4 ML algorithms, 5 real-world datasets and 1600 fairness evaluation cycles. We find a linear relationship between data and model fairness metrics when the distribution and the size of the training data changes. Our results indicate that testing for fairness prior to training can be a "cheap" and effective means of catching a biased data collection process early; detecting data drifts in production systems and minimising execution of full training cycles thus reducing development time and costs.

KEYWORDS

SE4ML, ML Fairness Testing, Empirical Software Engineering, Data-centric AI

ICSE-Companion '24, April 14–20, 2024, Lisbon, Portugal

ACM ISBN 979-8-4007-0502-1/24/04
https://doi.org/10.1145/3639478.3643121

ACM Reference Format:
Arumoy Shome, Luís Cruz, and Arie van Deursen. 2024. Data vs. Model Machine Learning Fairness Testing: An Empirical Study. In *2024 IEEE/ACM 46th International Conference on Software Engineering: Companion Proceedings (ICSE-Companion '24), April 14–20, 2024, Lisbon, Portugal.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3639478.3643121

1 INTRODUCTION

In contrast to prior work, we take a more holistic approach by testing for fairness at two distinct locations of the ML development lifecycle. First, prior to model training using fairness metrics that can quantify the bias in the training data (henceforth Data Fairness Metric or DFM). And second, after model training using fairness metrics that quantify the bias in the predictions of the trained model (henceforth Model Fairness Metric or MFM).

While MFM has been widely adopted in practice and well researched in academia, we do not yet know the role of DFM when testing for fairness in ML systems. The research goal of this study is to evaluate the effectiveness of DFM for catching fairness bugs. We do this by analysing the relationship between DFM and MFM through an extensive empirical study. The analysis is conducted using 2 fairness metrics, 4 ML algorithms, 5 real-world tabular datasets and 1600 fairness evaluation cycles. To the best of our knowledge, this is the first study which attempts to bridge this gap in scientific knowledge. Our results are exploratory and open several intriguing avenues of research.

The research questions are listed below. All source code and results of the study are publicly accessible under the CC-BY 4.0 license[1].

RQ1. **What is the relationship between DFM and MFM as the fairness properties of the underlying training dataset change?**

RQ2. **How does the training sample size affect the relationship between DFM and MFM?**

RQ3. **What is the relationship between DFM and MFM across various training and feature sample sizes?**

2 OUR APPROACH

We train 4 ML models on 8 datasets producing 32 total cases. The fairness for each case is evaluated 50 times using two fairness metrics, thus producing a total of 1600 training and fairness evaluation cycles.

A 75–25 split with shuffling is used to create the training and testing splits. DFMs and MFMs are used to quantify the bias in the underlying distribution of the training set and the predictions of the models respectively. We adopted the transformation steps from prior work to scale all fairness metric values between 0 and 1 such that higher values indicate more bias [4].

We further experiment with different number of examples and different number of features in the training set. For both experiments, we shuffle the order of the examples in the training and testing sets. Additionally, for the feature sample size experiment we shuffle the order of the features.

For the training sample size experiment, we generate different training samples of varying sizes starting from 10% of the original training data, and increase in steps of 10% until the full quantity is reached. For the feature sample size experiment, we start with a minimum of three features (in addition to the protected attribute and target) and introduce one new feature until all the features are utilised. Both the training and testing sets undergo the same feature sampling procedure in the feature sample size experiment.

We use correlation analysis to study the relationship between DFM and MFM with-respect-to change in three experimental factors—distribution, size and features of the training set. Spearman Rank Correlation is used to quantify the linear relationship between the DFM and MFM since it does not assume normality and is robust to outliers. We repeat all experiments 50 times and report the statistical significance of our results. We consider cases where $pvalue \leq 0.05$ to be statistically significant in our evaluation.

3 RESULTS & IMPLICATIONS

Results from RQ1 indicate that the DFM and MFM convey the same information when the distribution and consequently the fairness properties of the training data changes. ML systems running in a production environment are often monitored to detect degradation in model performance. As shown in Figure 1, a standard practice is to combine the data encountered by the model in the production environment with its predictions to create the training set for the next iteration [2]. Since data reflects the real world, change in its underlying distribution over time is eminent. Our results indicate that DFM can be used as a early warning system to identify fairness related data drifts in automated ML pipelines.

Results from RQ2 show that the quantity of training data significantly impacts the relationship between DFM and MFM . With smaller training sizes, there's a positive correlation between DFM and MFM, indicating bias in both training data and model predictions. As training data increases, this correlation diminishes, suggesting models learn to make fairer predictions despite biases in training data. However, *Zhang et al. (2021)* [4] highlight that data quality is also crucial; merely increasing data quantity doesn't necessarily resolve model biases. The study also notes a trade-off between model efficiency, performance, and fairness. Practitioners might reduce training data for efficiency, potentially impacting model performance and necessitating additional fairness mitigation efforts. Conversely, larger training datasets can reduce bias but require more computational resources. This balance between fairness, efficiency, and performance is a key consideration in ML system development and operation.

Results from RQ3 show a positive correlation between DFM and MFM as training sample size changes. This suggests that DFM can help practitioners identify fairness issues early, potentially saving the costs and energy associated with a full training cycle. Early detection of bias with DFM might also indicate problems in data collection or system design. However, this approach doesn't apply when altering the feature sample size of the training set, as larger feature samples usually enhance model fairness. Since no existing fairness metrics consider feature influence at the data level, it's advisable for ML practitioners to assess fairness both before and after training when experimenting with feature sample size.

4 CONCLUSION

This study introduces a novel approach to ML fairness testing by evaluating fairness both before and after training—using metrics to quantify bias in training data and model predictions, respectively. This "data-centric" approach is the first step towards integrating fairness testing into the ML development lifecycle. The study empirically analyzes the relationship between model-dependent and independent fairness metrics, finding a linear relationship when training data size and distribution change. The results suggest that testing for fairness before training an ML model is a cost-effective strategy for identifying fairness issues early in ML pipelines, and can aid practitioners navigate the complex landscape of fairness testing.

[1]The replication package of this study is available here: https://figshare.com/s/67206f7c219b12885a6f

REFERENCES

[1] Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann. 2019. Software engineering for machine learning: A case study. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. IEEE, 291–300.

[2] Felix Biessmann, Jacek Golebiowski, Tammo Rukat, Dustin Lange, and Philipp Schmidt. 2021. Automated Data Validation in Machine Learning Systems. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering.[Google Scholar]* (2021).

[3] Eric Breck, Neoklis Polyzotis, Sudip Roy, Steven Whang, and Martin Zinkevich. 2019. Data Validation for Machine Learning.. In *MLSys*.

[4] Jie M Zhang and Mark Harman. 2021. "Ignorance and Prejudice" in Software Fairness. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*. IEEE, 1436–1447.