CERIAS

The Center for Education and Research in Information Assurance and Security

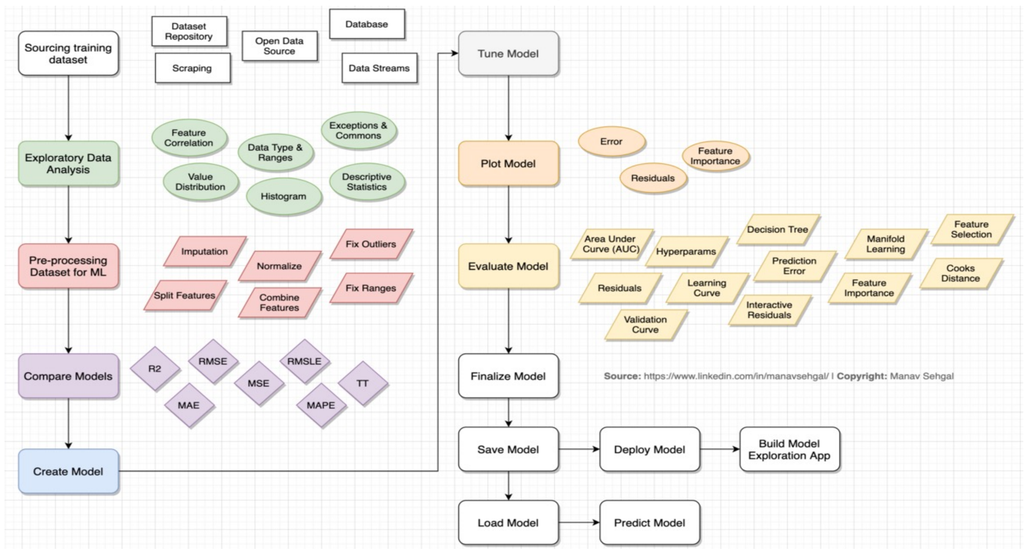
Estimating Machine Learning Model Fairness through Data Characteristics

Kevin Chittilapilly, Ahana Bhattacharya, Romila Pradhan

{kchittil, rpradhan}@purdue.edu, abhattacharyya@ubuffalo.edu

Motivation and Background

- As machine learning (ML) models find their way into various domains such as finance, healthcare, and law enforcement, there is growing concern that these models could unintentionally extend existing biases and discrimination present in the data they are trained on.
- Typical workflow: Acquire a dataset, select an ML model, tune model parameters and hyperparameters, evaluate performance



- To check the suitability of a dataset for a data science application, multiple models are trained on the dataset and checked for performance metrics
  - Computationally expensive, especially for complex ML models e.g., neural networks, which are resource-intensive
- Need to quickly estimate the fairness of models before they are deployed in real-world situations

We propose **DataFit**, a system that estimates the goodness of fit of a dataset for fairness of downstream ML tasks. We utilize advances in data profiling and meta-learning to estimate model fairness from historical data.

Our solution: DataFit

- Generate ***data characteristics or meta-features*** during a workflow
  - e.g., mean_skewness, pca_components_95_var, correlation, confidence
- Leverage past historical data to determine data characteristics that are indicative of specific fairness metrics by quantifying the metrics in terms of the data characteristics
- Learn a *meta*-model to estimate fairness given the data characteristics
- For a new dataset, use the learned meta-model to estimate fairness of an ML model trained over this dataset

Challenges: The new dataset obtained could be out of distribution from the training dataset. It might also be from a different sample leading to lower accuracy with learned models.

References

1. Galhotra, S., Fariha, A., Lourenço, R., Freire, J., Meliou, A., & Srivastava, D. (2021). DataExposer: exposing disconnect between data and systems. arXiv preprint arXiv:2105.06058.
2. Feurer, M., Springenberg, J., & Hutter, F. (2015, February). Initializing bayesian hyperparameter optimization via meta-learning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 29, No. 1).
3. Hofmann, H. (1994). Statlog (german credit data) data set. UCI Repository of Machine Learning Databases, 53.
4. Ding, F., Hardt, M., Miller, J., & Schmidt, L. (2021). Retiring adult: New datasets for fair machine learning. Advances in neural information processing systems, 34, 6478-6490.
5. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM computing surveys (CSUR), 54(6), 1-35.

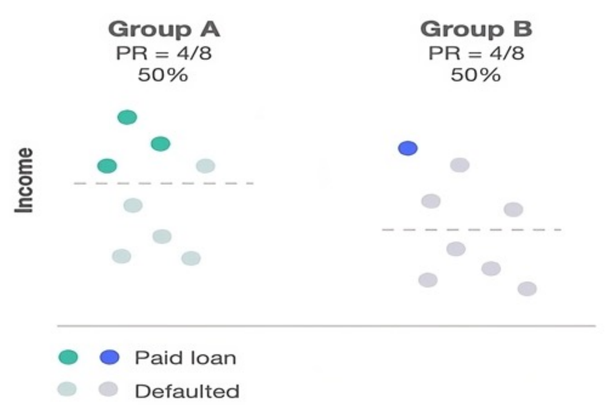
Experimental Setup

Datasets:

- German Credit dataset: contains demographic and financial information of 1,000 bank account holders. Prediction task: classify individuals as good or bad credit risk. [3]
- ACS Income dataset: contains demographic and employment information of 1,599,229 individuals. Prediction task: predict whether an individual's income is above $50,000. [4]

Performance metrics:

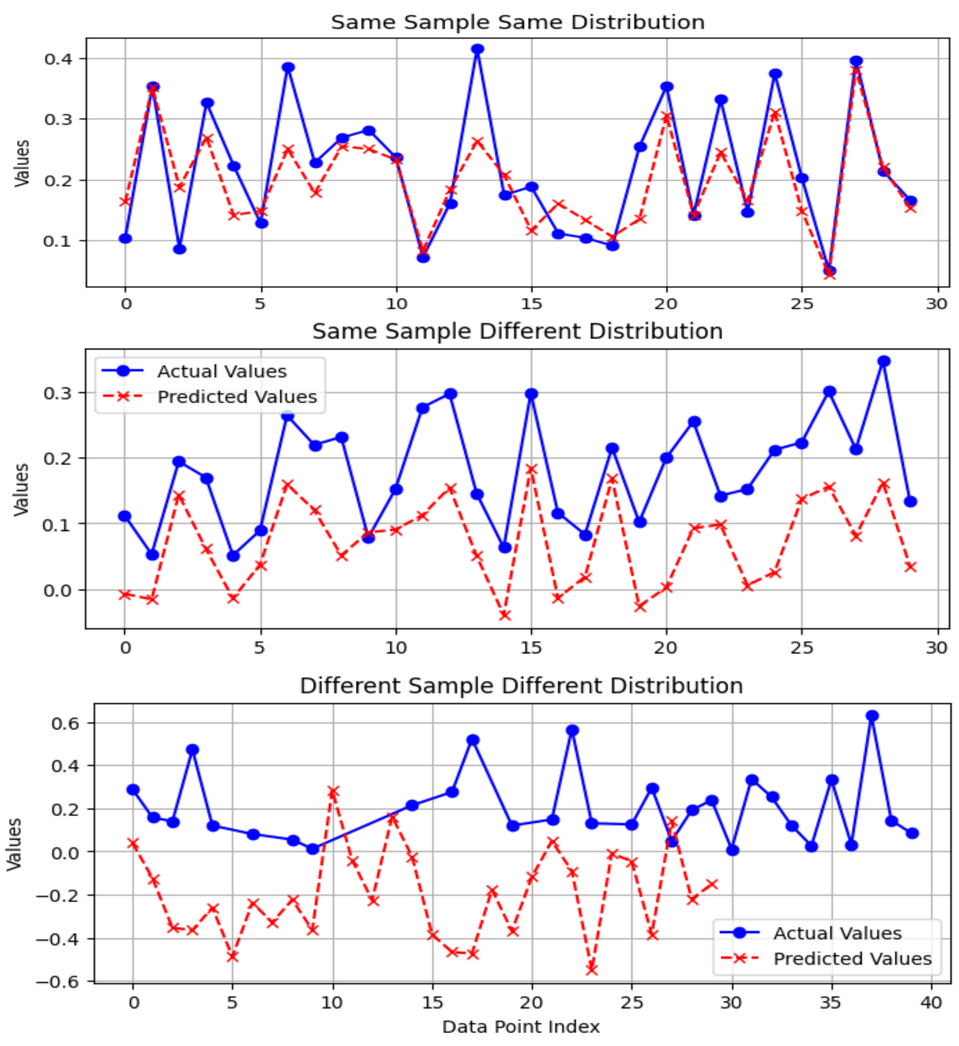
- Fairness metrics: Demographic parity, equalized odds, predictive parity. [5]
- e.g., an algorithm is considered fair according to demographic parity if different demographic groups have the same rate of positive (or negative) outcomes.



$$\Pr(\hat{Y} = 1 | A = 0) = \Pr(\hat{Y} = 1 | A = 1)$$

$\hat{Y}$ - Predicted values from model
A - Protected attribute (e.g., gender, race)

Experimental results



Conclusions

Estimating fairness of models trained over a dataset without actually training the model is of importance because of the huge time- and resource-intensive nature of the training process. Our results indicate that DataFit correctly estimates fairness for in-distribution and out-of-distribution data with huge runtime gains.

PURDUE UNIVERSITY

NSF

Google Research

CENTER FOR ADVANCING SAFETY OF MACHINE INTELLIGENCE
A collaboration between Northwestern University and UL Research Institutes

CERIAS