Dealing with Bias and Fairness in Data Science Systems: A Practical Hands-on Tutorial

Pedro Saleiro
Feedzai
pedro.saleiro@feedzai.com

Kit T. Rodolfa
Carnegie Mellon University
krodolfa@andrew.cmu.edu

Rayid Ghani
Carnegie Mellon University
rayid@cmu.edu

ABSTRACT

Tackling issues of bias and fairness when building and deploying data science systems has received increased attention from the research community in recent years, yet a lot of the research has focused on theoretical aspects and very limited set of application areas and data sets. There is a lack of 1) practical training materials, 2) methodologies, and 3) tools for researchers and developers working on real-world algorithmic decision making system to deal with issues of bias and fairness. Today, treating bias and fairness as primary metrics of interest, and building, selecting, and validating models using those metrics is not standard practice for data scientists. In this hands-on tutorial we will try to bridge the gap between research and practice, by deep diving into algorithmic fairness, from metrics and definitions to practical case studies, including bias audits using the Aequitas toolkit (http://github.com/dssg/aequitas). By the end of this hands-on tutorial, the audience will be familiar with bias mitigation frameworks and tools to help them making decisions during a project based on intervention and deployment contexts in which their system will be used.

KEYWORDS

Algorithmic Fairness, Bias Mitigation, AI Ethics

ACM Reference Format:
Pedro Saleiro, Kit T. Rodolfa, and Rayid Ghani. 2020. Dealing with Bias and Fairness in Data Science Systems: A Practical Hands-on Tutorial. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '20), August 23–27, 2020, Virtual Event, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3394486.3406708

1 MATERIALS

All the tutorial materials and instructions are available at https://github.com/dssg/fairness_tutorial

2 TARGET AUDIENCE AND PREREQUISITES

Data Scientists and practitioners from both the public and private sectors, and PhD students. Prerequisites: basics of classification models and evaluation methodologies.

KDD '20, August 23–27, 2020, Virtual Event, CA, USA

ACM ISBN 978-1-4503-7998-4/20/08.
https://doi.org/10.1145/3394486.3406708

3 IMPACT

There is a need for training data scientists and practitioners on how to deal with bias and fairness in practice, from the early stages of a data science project up to maintaining a ML system in production. Existing resources only cover the ML training/optimization part of bias mitigation and do not equip people with frameworks to help them making decisions during project execution on different contexts and application scenarios.

By the end of the tutorial, the audience will be equipped to take part in conversations around bias and helping decision makers understand the options and trade-offs involved; how to think about how different aspects of project scoping might influence fairness outcomes; defining what are the actions/interventions based on the model's prediction, what are the cohorts, target variables, evaluation metrics, bias and fairness goals for different groups; how to audit the model bias and fairness and continuous monitoring to assess the need for retraining.

4 TUTORIAL OUTLINE

(1) **Part 1**: Background, core concepts and discussion
  (a) Algorithmic decision-making and the role of data scientists.
  (b) Legal and regulatory aspects, public vs private sector conside-ations.
  (c) Understanding potential sources of bias
  (d) Bias and Fairness definitions
  (e) The Fairness Tree - or how to select appropriate metrics depending on actions/interventions
  (f) Bias mitigation approaches and techniques
    (i) Pre-modeling
    (ii) Model selection
    (iii) Regularization
    (iv) Thresholds
    (v) Applications
  (g) Additional Case Studies from Social Good Projects.
(2) **Part 2**: Hands-on bias audit, fairness-aware model selection and case studies
  (a) Intro to Case Study
  (b) Define evaluation mnetric, protected group, and bias metrics of interest
  (c) Aequitas toolkit overview
  (d) Bias Aware Modeling workflow using the Aequitas toolkit
    (i) Build a 'standard' model to predict risk of not getting funded
    (ii) Audit the model for fairness
    (iii) Build additional models
    (iv) Audit and look at possible tradeoffs to think about bias aware model selection

(e) Try other bias reduction methods
  (i) Pre-modeling: sampling, feature selection, etc.
  (ii) Regularization
  (iii) Group-specific thresholds
(f) Additional considerations and conclusions

(3) **Part 3**: Final Remarks
(a) Checklists
(b) Tools

5 TUTORS SHORT BIO

- **Pedro Saleiro** - Pedro Saleiro is a Data Science Manager at Feedzai where he leads the research group on AI Ethics. Previously, Pedro was a postdoc at the University of Chicago, working with Professor Rayid Ghani at the Center for Data Science and Public Policy, developing new methods, open source tools, such as Aequitas, and doing data science projects with government and non-profit partners in diverse policy areas. Pedro completed his PhD in Machine Learning and Information Retrieval at the Faculty of Engineering of the University of Porto.
- **Kit T. Rodolfa** - Kit Rodolfa is a Senior Research Scientist at Carnegie Mellon University, working with Professor Rayid Ghani at the intersection of machine learning and public policy on using these methods to benefit society. His research interests include the bias, fairness, and interpretability of machine learning methods. Previous, Kit led the initial data science efforts at Devoted Health, and has served as Chief Data Scientist at Hillary for America, as the Director of Digital Analytics for the White House Office of Digital Strategy during the Obama administration, and on the analytics team during President Obama's 2012 re-election campaign. He holds a PhD in Biology and Master's in Public Policy for Harvard University, MPhil in Chemistry from the University of Cambridge, and BS degrees in Physics and Chemistry from Harvey Mudd College.
- **Rayid Ghani** - Rayid Ghani is a Distinguished Career Professor in the Machine Learning Department and the Heinz College of Information Systems and Public Policy at Carnegie Mellon University. Rayid works on the use of large-scale AI/Machine Learning/Data Science in solving large public policy and social challenges in a fair and equitable manner. Among other areas, Rayid works with governments and non-profits in policy areas such as health, criminal justice, education, public safety, economic development, and urban infrastructure. Rayid is also passionate about teaching practical data science and started the Data Science for Social Good Fellowship that trains computer scientists, statisticians, and social scientists from around the world to work on data science problems with social impact.

REFERENCES

[1] Kit T. Rodolfa, Pedro Saleiro, and Rayid Ghani. Chapter 11: Bias and fairness. In Ian Foster, Rayid Ghani, Ron S Jarmin, Frauke Kreuter, and Julia Lane, editors, *Big data and social science: A practical guide to methods and tools*. crc Press, 2020.

[2] Kit T Rodolfa, Erika Salomon, Lauren Haynes, Iván Higuera Mendieta, Jamie Larson, and Rayid Ghani. Case study: predictive fairness to reduce misdemeanor recidivism through social service interventions. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 142–153, 2020.

[3] 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. (2018), nov 2018.

[4] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of Opportunity in Supervised Learning. *Advances in Neural Information Processing Systems*, (Nips):1–22, 2016.

[5] Solon Barocas and Andrew D. Selbst. Big Data's Disparate Impact. *California Law Review*, 104(3):671–732, 2016.

[6] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions. nov 2018.

[7] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*, pages 134–148, 2018.

[8] Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 2018.

[9] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On Fairness and Calibration. In I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5680–5689. Curran Associates, Inc., 2017.

[10] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness constraints: Mechanisms for fair classification. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017*, 54, 2017.

[11] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. *26th International World Wide Web Conference, WWW 2017*, pages 1171–1180, 2017.

[12] Andrew Cotter, Maya Gupta, Heinrich Jiang, Nathan Srebro, Karthik Sridharan, Serena Wang, Blake Woodworth, and Seungil You. Training Well-Generalizing Classifiers for Fairness Metrics and Other Data-Dependent Constraints. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 1397–1405, Long Beach, California, USA, jun 2019. PMLR.

[13] Indrė Žliobaitė. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, 31(4):1060–1089, jul 2017.

[14] Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness - FairWare '18*, pages 1–7, New York, New York, USA, 2018. ACM Press.

[15] Alexandra Chouldechova. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, 5(2):153–163, jun 2017.

[16] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhill Mullainathan. Human Decisions and Machine Predictions*. *The Quarterly Journal of Economics*, 133(January):237–293, aug 2017.

[17] Chris Russell, Matt J Kusner, Joshua Loftus, and Ricardo Silva. When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness. In I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6414–6423. Curran Associates, Inc., 2017.

[18] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual Fairness. In I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4066–4076. Curran Associates, Inc., 2017.

[19] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica, May*, 23:2016, 2016.

[20] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*, pages 259–268, New York, New York, USA, 2015. ACM Press.

[21] Alekh Agarwal, Aliiia Beygelzimer, Miroslav Dudfk, John Langford, and Wallach Hanna. A reductions approach to fair classification. *35th International Conference on Machine Learning, ICML 2018*, 1:102–119, 2018.

[22] Yahav Bechavod and Katrina Ligett. Penalizing Unfairness in Binary Classification. jun 2017.

[23] Naman Goel, Mohammad Yaghini, and Boi Faltings. Non-Discriminatory Machine Learning through Convex Fairness Criteria. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society - AIES '18*, pages 116–116, New York, New York, USA, 2018. ACM Press.

[24] Blake Woodworth, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. Learning Non-Discriminatory Predictors. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65, pages 1920–1953, Amsterdam, Netherlands, jul 2017. PMLR.