CERTIFAI: A Common Framework to Provide Explanations and Analyse the Fairness and Robustness of Black-box Models

Shubham Sharma
University of Texas at Austin
Austin, Texas
shubham_sharma@utexas,edu

Jette Henderson
CognitiveScale
Austin, Texas
jhenderson@cognitivescale.com

Joydeep Ghosh
CognitiveScale
Austin, Texas
jghosh@cognitivescale.com

ABSTRACT

Concerns within the machine learning community and external pressures from regulators over the vulnerabilities of machine learning algorithms have spurred on the fields of explainability, robustness, and fairness. Often, issues in explainability, robustness, and fairness are confined to their specific sub-fields and few tools exist for model developers to use to simultaneously build their modeling pipelines in a transparent, accountable, and fair way. This can lead

ACM Reference Format:
Shubham Sharma, Jette Henderson, and Joydeep Ghosh. 2020. CERTIFAI: A Common Framework to Provide Explanations and Analyse the Fairness and Robustness of Black-box Models. In *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES '20), February 7–8, 2020, New York, NY, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3375627.3375812