Thesis Project Portfolio

Equitable Artificial Intelligence: A Guide for Meta-Analysis of Techniques for De-Biasing Machine Learning Models

(Technical Report)

Examining Bias in Machine Learning Models of Financial Institutions

(STS Research Paper)

An Undergraduate Thesis

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Sophie Meyer

Spring, 2023

Department of Computer Science

Contents of Portfolio

Executive Summary

Machine learning has become an increasingly popular technology, allowing computers to learn and make decisions based on a given data set. It can be applied to various areas, including image classification, chatbots, GPS routing, and loan approvals. In loan approvals, machine learning is used to determine who should receive a housing loan, with the computer being trained with data from past loan applicants to facilitate this decision-making. However, AI systems are not inherently neutral. They are developed by humans who may inadvertently or deliberately incorporate their biases into the system. These biases can be amplified by machine learning models and lead to discriminatory outcomes. For example, the use of machine learning models to determine loan eligibility can perpetuate housing loan discrimination, as described in my STS research paper. Therefore, it is crucial that the developers of AI systems take into account the potential for bias and work towards developing systems that are unbiased, ethical, and socially responsible. To do this, existing methods of reducing bias in these systems can be examined. My technical report functions as a guide for meta-analysis of techniques for de-biasing machine learning models.

In the United States of America, there exists a long-standing history of housing discrimination which plagues both our nation's past and, unfortunately, present. In the early 20th century, discriminatory lending practices known as redlining denied mortgages to Black Americans in certain neighborhoods, perpetuating segregation and limiting opportunities for building wealth. The emergence of machine learning algorithms has the potential to perpetuate this discrimination. Recent studies have shed light on the potential for bias in AI systems that determine loan eligibility. For instance, a 2019 study analyzed data from Fannie Mae and Freddie Mac, including data about mortgage applications reviewed by both face-to-face lenders and AI algorithms. The study found that although FinTech algorithms discriminate 40% less compared to face-to-face lenders, they still exhibit biases. Specifically, Latinx and Black borrowers on FinTech platforms pay higher interest rates than their white counterparts. This finding indicates that although AI systems may reduce bias, they are not immune to it.

To determine the best existing method or combination of methods for debiasing machine learning models, a systematic evaluation and meta-analysis of the data is needed. This can aid socially responsible use of machine learning technology. Performing a meta-analysis of debiasing techniques involves several steps, outlined in the technical report. First, identifying reputable studies and techniques outlined in them, which are then sorted according to their applicability. As not all debiasing methods work for all applications, techniques must be categorized based on the type of algorithm they can be used on and the type of task they can be used for. After categorizing, tests must be developed for each category and criteria must be established for what makes a “good” or “bad” algorithm. Such a criteria may look at correlations between outcomes and demographics. Finally, tests must be performed, and results gathered and analyzed to determine the best existing debiasing techniques.

Future work includes performing the de-biasing meta-analysis. The results from this research can be used to guide further exploration and development of the techniques that are found to be most promising. Additionally, as the field of machine learning is constantly evolving, new examples of bias in machine learning will appear, and those must be examined and mitigated with similar analyses. As for the specific case of bias of banks, a full survey of the problem would be helpful. One step in reducing bias in housing loans is to better understand the problem. To do so requires loan data from banks. This is difficult to do now, as many banks would like to ensure that their information is kept private. Allowing researchers to access loan data would likely open them up to lawsuits and could give their competitors an advantage. However, if banks were offered an opportunity to anonymously contribute to research, perhaps further research could be done. Once the problem has been outlined, de-biasing methods could be selected to test. By doing so, we can ensure that AI technology is used to promote equality and social justice, rather than perpetuate discrimination and inequality.