A Meta-Analysis of Techniques for De-Biasing Machine Learning Models

Examining Bias in Machine Learning Models for Housing Loan Approvals

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By
Sophie Meyer

October 27, 2022

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISORS

Kent Wayland, Department of Engineering and Society

Briana Morrison, Department of Computer Science

General Research Problem: Examining and Reducing Bias in Machine Learning Models

How does bias become part of and manifest in machine learning, and how can this bias be reduced?

Machine learning is a technique in which a computer is given data and produces a model by "learning" about the data. (Zhou, 2021). For example, one could feed in pictures of dogs and a label of its breed, and the model that is produced would be able to identify the breed of dog.

Machine learning is often perceived as an objective method of evaluation, but in actuality machine learning models can share the same biases as the humans who create them. According to mathematician, data scientist, and former hedge fund trader Cathy O'Neil in her book *Weapons of Math Destruction*, "Models are opinions embedded in mathematics" (O'Neil, p. 8). This sentiment becomes extremely apparent when examining the machine learning models involved in loan approval, where racial biases affect the models irrespective of other factors. In my STS paper, I will discuss the role that bias plays in loan approval machine learning models, and in my technical paper, I will address possible methods to mitigate such bias. The overarching theme of bias in machine learning models connects these two topics.

Examining Bias in Machine Learning Models for Housing Loan Approvals

What Role does Historical and Current-Day Racial Bias Play in Machine Learning Models used for Housing Loan Approval?

O'Neil's *Weapons of Math Destruction* details a study wherein a disturbing fact was uncovered: a certain unnamed bank's machine learning model involved in loan approvals was less likely to do so from zip codes that have higher percentages of Black residents. This was controlling for income, credit score, and all other outside factors. Effectively, race alone had become a deterministic factor in loan approval. Even though it is technically illegal to discriminate against others based upon race, the bank itself was neither directly nor intentionally doing so, rather the machine learning model had learned to discriminate based on proxies. That is, the model had learned that certain aspects, like zip code, tend to correspond with people of certain races, and therefore was still able to produce biased results despite not being explicitly told race.

In *The Case for Reparations,* Ta-Nehisi Coates outlines the ways in which Black people have been systematically discriminated against, and how said discrimination has impacted their struggles for modern-day financial liberation (Coates, 2014). He advocates for a reckoning with "our compounding moral debts." Throughout American history, Black people have been oppressed through both legal and illegal means. Even after slavery, predatory loaning practices and redlining made it nearly impossible for Black Americans to participate in the home-mortgage market, and thereby depriving them of the means by which many white Americans began to build and maintain generational wealth. According to Coates' article, even in the present day, the data show that Black people earning upper-middle class incomes still do not live in the same kinds of neighborhoods that white people with equivalent incomes do. This connects to the section in *Weapons of Math Destruction* (O'Neil, 2016) that outlines how loan decisions are made based on zip codes. If Black people are living in different neighborhoods, as this article states, then it is especially easy for the model to pick out which neighborhoods are majority

Black. The article poses the question of whether it is fair to determine loan eligibility based on income and people's financial past, given the inequalities that were outlined, and those that persist today.

The Social Construction of Technology (SCOT) theory posits that social factors influence the development and use of technology (Mitcham, 2005). This theory uses the ideas of relevant social groups, which are the people, organizations, and groups that influence and are influenced by a particular technology. Additionally, SCOT employs the concept of interpretive flexibility, the idea that technological artifacts are "open to radically different interpretations by various social groups" (Mitcham, 2005, p. 1791).

Drawing on SCOT, my STS paper will explore the ways in which the history of housing loans has impacted the machine learning loan approval models of today. Technological determinism is a term for the view that technology drives the course of society. This concept of technological determinism is not sufficient to explain the interaction between loan decisions and the machine learning models that increasingly determine them. Of course, the technology now has an effect, but the social roles that led to the creation of specific models must also be explored. In order to draw upon SCOT, the paper will delve into interactions between the relevant social groups, including banks that offer housing loans and Black people looking for housing that are negatively affected by loan denials. The idea of interpretive flexibility can also be applied, housing loans have been, and continue to be, interpreted differently by different groups. People who have been denied loans may see them in an entirely different light than, for example, bank employees and owners, who may see them as merely a way to make money, or people who have been approved for loans. The development and application of concepts from

SCOT will address the question of how the social factor of racial bias has helped to construct machine learning models for housing loans.

To collect data to explore this question, I will conduct a case study of Bank of America, which uses artificial intelligence in part to determine loan decisions. It also currently has a program called the Community Affordable Loan Solution™ that aims to expand homeownership opportunities in Black and Hispanic-Latino communities (Konish, 2022). Because Bank of America has, in recent years, been instituting policies that have the goal of reducing inequality, it is a good candidate for a case study. The case study will be placed into the context of the background research, and analyzed through the lens of SCOT.

A Meta-Analysis of Techniques for De-Biasing Machine Learning Models

What approach is best for debiasing machine learning datasets and models? Biased outputs based on machine learning can have real and devastating effects on people. One example of this is outlined in my STS paper: the issue of housing loans being denied based on race. Another example is described in a 2017 study looking at accountability mechanisms and legal standards for algorithms. The researchers found that the algorithm used to determine who would receive a visa through the State Department's diversity visa lottery was biased, meaning that people would not be able to have a fair chance at receiving a visa to immigrate to the United States (Kroll, 2017). These are only two of the many issues raised by this topic, and a good solution would be invaluable.

There exists a large number of proposed methods for debiasing, focusing on everything from data collection through deployment of the machine learning model. A systematic evaluation and meta-analysis of the data will be used in order to determine the best existing method. The

results of each examined method will be examined to see which performed best in their way of quantifying bias. It is important to note that it is difficult to directly compare amount of bias, because each de-biasing method is tested on a different model with different datasets, and therefore this examination will need to be, at least in part, qualitative (Martín, 2022). A method that works for one type of model may not work for others, but using a qualitative approach will allow for determining what types of cases may be best de-biased by certain methods.

The existing research that is most relevant to my STS paper is in an article that examines the accuracy of machine learning models in predicting loan defaulting (Fu et. al, 2021). It finds that, while the models that the researchers looked at were highly accurate, they also showed signs of gender and racial bias. The models were not directly given gender and race as inputs, but still demonstrated bias. The study presented a de-biasing method for the model, which was also tested for prediction accuracy. It was found to be a slightly less accurate predictor than the biased model, but still a stronger predictor than the human participants in the study.

The first place when creating a machine learning model that bias can be introduced is in the data that is fed in to train the model. This idea of "garbage in, garbage out" means that if you used data that is biased to train the model, then the model will learn to be biased (DeBrusk, 2018). This has been shown to be the cause of many biased models, such as a certain infamous one trained by Microsoft. It was supposed to generate tweets using machine learning, and very quickly turned to "praising Hitler and spewing out misogynistic remarks" (DeBrusk, 2018). Another example is the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), which was an attempt at creating a model to determine whether a previous offender was likely to reoffend. It however was more likely to predict that a Black person was more likely to reoffend than a white person with the same history. The cause in this case was also shown to

be biased data, where the training data included race, but did not include many other pieces of data like details of past arrests. When the data was corrected to not include race and include many other factors, the bias was shown to go away.

However, a 2021 study pushes back on the common refrain that bias is introduced solely or primarily through biased data (Suresh, et al.) Rather, researchers find that bias may come from more than one place in the process of building, using different mathematical methods, and testing the model. Firstly, they mention that biased data is not simply an issue of sampling incorrectly, but that the process of obtaining the data is "long and complex, grounded in historical context and driven by human choices and norms." Second, they mention that human decisions are made at every point in the design cycle of a model, and any of these can introduce bias. The article seeks to identify the sources of harm in these models. My technical paper will seek to examine each stage in the machine learning lifecycle and perform a meta-analysis on the various technologies that have been developed to prevent and mitigate this bias. A study from the University of Pennsylvania explores "how…computational techniques can assure that automated decisions preserve fidelity to substantive legal and policy choices" (Kroll, 2017, p. 1). They "show how these tools may be used to assure that certain kinds of unjust discrimination are avoided and that automated decision processes behave in ways that comport with the social or legal standards that govern the decision" (Kroll, 2017, pp. 1-2).

An additional study proposes the idea of using biologically based approach, incorporating randomness, to de-bias data. The study shows very strong results, and therefore will be heavily considered in the meta-analysis. (Liu, et. al, 2016).

Conclusion

Machine learning models are capable of reflecting the bias of their human creators. The racial bias of the past and present has lead to the biased technology of loan models today. And, in the interest of equity, there have been many efforts to reduce or even eliminate the bias in models that are not fair. A meta-analysis of some such efforts will be performed to analyze which of these attempts are most effective. In the future, this research could be used to ensure that no one is unfairly denied a loan.

Citations

Castro Martín, L. (2022). Computational methods for bias reduction in surveys.

Coates, T.-N. (2014, May 22). The Case for Reparations. The Atlantic. https://www.theatlantic.com/magazine/archive/2014/06/the-case-for-reparations/361631/

DeBrusk, C. (2018). The risk of machine-learning bias (and how to prevent it). *MIT Sloan Management Review*.

Fu, R., Huang, Y., & Singh, P. V. (2021). Crowds, Lending, Machine, and Bias. *Information Systems Research*, 32(1), 72–92. https://doi.org/10.1287/isre.2020.0990

Johnson, D. (2005). Social construction of technology. In C. Mitcham (Ed.), *Encyclopedia of Science, Technology, and Ethics (Vol. 4, pp. 1791–1795)*. Macmillan Reference.

Konish, L. (n.d.). Bank of America launches zero down payment mortgages to help minorities buy their first homes—Here's who can apply. CNBC. Retrieved December 2, 2022, from https://www.cnbc.com/2022/09/01/bank-of-america-to-help-minorities-buy-first-homes-with-new-mortgages.html

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable Algorithms. *University of Pennsylvania Law Review, 165(3), 633–705.* http://www.jstor.org/stable/26600576

H. Liu, A. Gegov and M. Cocea, "Nature and biology inspired approach of classification towards reduction of bias in machine learning," *2016 International Conference on Machine Learning and Cybernetics (ICMLC)*, 2016, pp. 588-593, doi: 10.1109/ICMLC.2016.7872953.

O'Neil, C. (2016). *Weapons of Math Destruction*. Penguin Books.

Suresh, H., & Guttag, J. V. (2021). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. *ArXiv*. http://arxiv.org/abs/1901.10002

Zhou, Z. H. (2021). Machine learning. *Springer Nature*.