Examining Bias in Machine Learning Models of Financial Institutions

A Research Paper submitted to the Department of Engineering and Society

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

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Sophie Meyer

Spring 2023

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

Advisor

Kent Wayland, Department of Engineering and Society

Introduction

Having become increasingly popular in recent years as our dependence upon technology grows ever more apparent and involved in our ongoing day-to-day activities, the term “machine learning” is used to describe a process whereby a computer is trained to learn and make decisions based on a given dataset (Zhou, 2021). Once fed data, the computer crafts a model that can be used for decision-making. Machine learning has a plethora of applications, ranging from image classification to chatbots, GPS routing to loan approvals. Its capabilities are incredibly diverse, capable of determining the breed of a dog based off an image, facilitate faster navigation to a destination, or enable chatbots to provide prompt responses to user queries. In the context of 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. With machine learning, the computer carefully considers relevant factors such as income and credit score. But just because computers are not human does not make them immune to some of the more human flaws which infest our society, including biases. In this paper, we will first address the history of both racial and gender bias in the American financial sector, and then examine how-- if at all-- machine learning models have been, continue to be, and can be in the future used to rectify such biases.

Background

The biases which we refer to in this paper are not to be confused with the mathematical and statistical definition of bias commonly used in the field of machine learning, wherein the proximity between estimated and actual values is calculated in order to ascertain how well a model is trained using sample data. Rather, when we speak of “bias” we are referring to the human capability of discrimination, whether intentional or unintentional, conscious,

subconscious, or unconscious. Such biases manifest in many different forms, some innocuous, all insidious. When this version of "bias" is discussed, one's mind likely turns to the more prevalent ones which pervade society, such as racism and sexism, which are the systemic discrimination of others on the basis of race and sex, respectively.

In the United States of America, there exists a long-standing history of bias which plagues both our nation's past and, unfortunately, present. This bias is often communicated by and controlled via financial means, no more clear than during the 246 years—from 1619 to 1865, to be exact—during which the state permitted people to be considered property. 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. He advocates for a reckoning with "our compounding moral debts." Even long after the Civil War, the relationship between race, finance, and bias remained cemented, as evidenced through the practices of redlining, wherein individuals residing in predominantly Black neighborhoods were systematically denied credit, insurance, loans, and other financial services, and reverse-redlining, wherein majority minority neighborhoods were targeted with inflated interest rates. Such predatory lending practices persisted until the late 60's up until the Fair Housing Act of 1968 was passed, making 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 shows that Black people earning upper-middle class incomes still do not live in the same kinds of neighborhoods that white people earning equivalent incomes do. *Weapons of Math Destruction* outlines how loan decisions are often made based on zip codes (O'Neil, 2016). If Black people are living in different neighborhoods,

as Coates states in his article, then it is especially easy for the model to guess an individual's race based on which neighborhoods are majority Black. Cathy O'Neil's novel *Weapons of Math Destruction* details a study wherein a disturbing fact was uncovered: a 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 after 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.

Women, as well, have been systematically excluded from many areas of life that are most crucial in building , generating, and maintaining wealth. Until 1848, women could not own property, sign contracts, or control their own income in the United States. Even when women were able to own property, they faced significant challenges in accessing credit and financial resources independently. For example, until the Equal Credit Opportunity Act of 1974 was passed, women could not obtain credit cards in their own name. This meant that they were often dependent on their husbands or fathers to provide them with access to credit, making it difficult for them to build a credit history and thereby establish financial independence (Sandberg).

Many turn to machine learning in hopes that the addition of more advanced technology to the financial sector will help remedy such biases by minimizing the role which face-to-face lenders and creditors play in gatekeeping financial independence from the sizeable, disenfranchised portion of the population of women, racial minorities, and of course female

minorities. However, while often perceived as a more objective method of evaluation, machine learning models are not immune to the biases of the humans who help create them. Rather, they often end up inheriting and propagating such biases.

In her book *Weapons of Math Destruction,* mathematician, data scientist, and former hedge fund trader Cathy O'Neil claims that "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, and those involved in determining credit limits, where gender biases affect the models disproportionately as well.

In March of 2021, the Berkeley Haas Center for Equity, Gender, and Leadership on Mitigating Bias In Artificial Intelligence published their study wherein they analyzed over a hundred biased AI systems, such as facial recognition systems, across various industries over the past thirty years. The center found that 44.2% (59 systems) demonstrated gender bias. What's more, 25.7% (34 systems) were identified exhibiting both racial and gender bias. In one example, a husband and wife found their Apple Card credit limits had a 20x difference, with the woman having an exponentially lower financial resources at her disposal, all because the algorithm deemed her less creditworthy than the man she was married to.

In her 2018 article "Gender Bias in Artificial Intelligence", USC information communications expert Susan Leavy speaks to the reduction of gender bias in machine learning. She cites decades of over-representation of men in the field of computer science as one source of creator bias creeping into machine learning models. However, this is just one facet of the problem. Leavy's article also mentions the fact that centuries of sexism means that misogynistic ideas and language pervades all training data, even without the creator's intention. She argues

that society impacts technology, and therefore in order to reduce bias, we must consult critical theory and look to analysis of gender theory in order to create the most unbiased data possible. We must try to create an AI that lacks the bias that is so ubiquitous in our society, making the AI better in fact than us humans. Leavy concludes that current debiasing methods are inadequate, and that both a historical perspective ought to be explored and past biases must be identified in order to ethically create machine learning models which adequately address gender disparities.

Each year, JP Morgan Chase, an international banking corporation with a large-scale client base, publishes its annual report. Since 2016, each of their annual reports have explicitly mentioned either machine learning and/or artificial intelligence. In their 2021 report-- the most recent made publicly available—JP Morgan mentions that they uses artificial intelligence to "generate insights on existing and prospective clients from public information". They claim that their machine learning models are used in conjunction with customer data to make predictions and gain further insights. This is a clear example of a bank that is not only already heavily investing in this technology but also which promises to continue to improve their existing systems. Since 2017, they have invested over $100 million on artificial intelligence, machine learning, and other "technology initiatives" targeted solely for the reduction of fraud. Overall, though they decline providing hard figures, JP Morgan claims they have spent upwards of "hundreds of millions of dollars" on their various AI initiatives. To their credit, they have had no scandals related to the use of AI, and as of yet there is no publicly available data that suggests that their use of AI thus far is biased. What's more, JP Morgan claims to have "deployed or committed more than $18 billion of $30 billion to advance racial equity." While it is unclear what exactly this $18 billion has been spent on, given their recent investment portfolio, it is safe

to say that some of this fund has been allotted to AI debiasing, whose role is of great importance in bringing us closer to racial equity, especially in the field of finance.

In her 2021 study published with the title “Crowds, Lending, Machine, and Bias”, NYU economist Runshan Fu describes machine learning as an alternative to crowd lending. Crowd lending is a means by which businesses raise without relying on the services usually provided by a bank. Instead, crowds of investors personally choose which businesses to loan money to. Fu’s study found that when a machine learning algorithm was implemented and tested with regard to crowd lending, it outperformed crowds in predicting which businesses would ultimately default on their loans. Additionally, using the AI model led to a higher rate of return for lenders and more “funding opportunities for borrowers with few alternative funding options” (p. 89). However, the researchers found that the model was still biased, even though it was not told to use race or gender as parameters. The researchers attributed this failure to omit bias from the model to redundant encodings, that is, features that the model used as proxies for gender and race in the absence of these protected attributes. Some examples of redundant encodings for race could be name origin or languages spoken. In response, they then implemented a debiasing algorithm to remove these redundant encodings so that race and gender could not be inferred. Overall, their debiased model had only a small reduction in accuracy for the test set, around 2%.

Methods

The sources referenced in this paper are mostly scientific and/or scholarly in nature, with most falling into one of three broad categories. The first category of sources is those which provide evidence of existing bias in machine learning models. These sources help develop the background necessary for identifying the root of the issue when it comes to bias in machine learning. The second category of sources include articles which describe the different types of

bias encountered in machine learning. Many note the success of different de-biasing efforts which have been made that, as the name suggests, aim to counteract technological discrimination against women and minorities. The third category of sources are statements and articles from banks and other entities which create and deploy these machine learning models. In the past few years, there has been an increased awareness of the bias that pervades machine learning. In response, these institutions claim to have been and currently be making efforts to mitigate this bias, though the merit of such claims is up for debate.

Findings

There is a long record of loan discrimination against African Americans in particular. From the times of slavery, when Black people were unable to own property, through the Jim Crow era and beyond, Black communities have faced systematic discrimination in lending practices. Women, too, were barred from independent financial participation for much of the nation's history. Thus, there is a demonstrated history of bias against minority groups at the hands of human lenders, creditors, bankers, and other individuals with financial power and influence.

Meanwhile, there is significantly less data and, of course, history behind racial and sexual bias in banking done by machines. Banks do not tend to be very forthcoming with data about who they offer loans to, and certainly are not letting outside parties have access to or test their algorithms. However, we do know that more banks are using AI, and will continue to integrate it even more into their workings as it becomes more technologically advanced. Many papers describe how easily bias is encoded into algorithms, and how it must be consciously avoided.

A 2019 paper by Bartlett et. al analyzed data from Fannie Mae and Freddie Mac, including data about mortgage applications reviewed by both face-to-face lenders and AI algorithms. It found that although compared to face-to-face lenders, FinTech algorithms discriminate 40% less, on FinTech platforms, Latinx and African-American borrowers pay slightly higher interest rates - 5.3 basis points more for purchase mortgages and 2.0 basis points more for refinance mortgages.

There have already been numerous attempts made to reduce bias, with a diverse range of approaches. One example can be found in ZestFinance—recently rebranded as ZestAI-- a company committed to create underwriting equity through the help of machine learning. Their founder and CEO Douglas Merrill believes that AI can and should be less biased than humans. ZestFinance offers a tool, ZAML Fair, that ranks types of data, such as zip code or age, by how likely they are to produce a biased outcome. Types of data that are more likely to produce biased outcomes are then weighted less heavily. ZestFinance claims that "70% of the mortgage approval rate gap between Hispanic and white borrowers" could be removed with their technology, amounting to over 100 million homeowners that would previously be denied loans (Fuscaldo, 2023). There are many AI companies that offer services to banks, but most do not go as far to address inequalities. ZestFinance was an optimistic look at what can be done when technology is created in a socially responsible way. However, in 2020 it settled a class action lawsuit for offering payday loans with interest rates far over the legal limit (Titus, 2018). While the ideas behind ZestFinance have potential, the company instead preyed upon vulnerable people. To this day, they continue to work with many large institutions, all seemingly unfazed by the company's less-than-stellar record.

Another example can be found in a 2019 paper published by the National Bureau of Economic Research entitled "Consumer-Lending Discrimination in the FinTech Era". In this study, a research committee consisting of lawyers and economists from UC Berkeley compared the efficacy of bias reduction in the machine learning models of the financial technology industry with face-to-face human lenders. They found that machine learning algorithms "do indeed remove some face-to-face biases". In fact, they found that these algorithms— which they refer to as "FinTechs" for short— "discriminate 40% less on average than face-to-face lenders" (p. 1). Meanwhile, face-to-face lenders were found to charge Latinx and African-American borrowers 7.9 and 3.6 basis points more for purchase and refinance mortgages respectively, costing them $765M in aggregate per year in extra interest" (p. 2).

This isn't to say that the machine learning models are perfect or even without flaws, however. They note that there remains a "potential for inaccuracies in [their] estimation in discrimination due to errors in identifying borrower race or ethnicity" (p. 48).

Conclusion

Housing loan discrimination in the US has a long history, and the use of machine learning models to determine loan eligibility has the potential to perpetuate this discrimination. The mere fact that JP Morgan spends millions of dollars on reducing bias in artificial intelligence is proof that this will be an issue with many algorithms. And, although many institutions claim to be reducing this bias, it is possible that something is slipping through the cracks. As O'Neil pointed out, there is a considerable amount of opacity related to these machine learning models that determine things as impactful as whether someone can get a loan.

But there is an important question to address—one of extreme importance to the banks and hedge funds themselves, which are often more concerned with profit made over people impacted: what if profit is higher with bias?

Certain machine learning models that incorporate bias net a higher profit for the company. So does this justify keeping bias in these models if it turns a higher profit?

Because current discrepancies in wealth and income arise from past discrimination, it is imperative that this discrimination not be perpetuated using artificial intelligence. Even if banks would make more money by refusing loans to more Black people, they should not. Perpetuating discrimination is not only morally wrong but also counterproductive in the long run. By denying loans to qualified individuals based on their race or ethnicity, banks are effectively limiting their potential customer base and, therefore, their profits.

Therefore, it is crucial that we develop and use AI systems in a way that is ethical, unbiased, and socially responsible, and that takes into account the historical and societal context in which they operate. By doing so, we can ensure that AI technology is used to promote equality and social justice, rather than perpetuate discrimination and inequality.

References

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

Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2019). Consumer-Lending Discrimination in the FinTech Era (Working Paper No. 25943). National Bureau of Economic Research. https://doi.org/10.3386/w25943

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/

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

Titus v. Zestfinance, Inc., CASE NO. 18-5373 RJB | Casetext Search + Citator. (n.d.). Retrieved April 25, 2023, from https://casetext.com/case/titus-v-zestfinance-inc

Leavy, S. (2018). Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. Proceedings of the 1st International Workshop on Gender Equality in Software Engineering, 14–16. https://doi.org/10.1145/3195570.3195580

Zhou, Z.-H. (2021). Machine Learning. Springer Nature.

Fuscaldo, D. (n.d.). ZestFinance Using AI To Bring Fairness To Mortgage Lending. Forbes. Retrieved May 16, 2023, from https://www.forbes.com/sites/donnafuscaldo/2019/03/19/zestfinance-using-ai-to-bring-fairness-to-mortgage-lending/

Sandberg, E. (n.d.). The History Of Women And Credit Cards. Bankrate. Retrieved May 16, 2023, from https://www.bankrate.com/finance/credit-cards/history-of-women-and-credit-cards/

Mitigating Bias in Artificial Intelligence. (n.d.). Berkeley Haas. Retrieved May 16, 2023, from https://haas.berkeley.edu/equity/industry/playbooks/mitigating-bias-in-ai/