The Effect of Data-based Economic Metrics on Marginalized Identity Groups

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

Khushi Chawla

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

Travis Elliot, Department of Engineering and Society

Introduction

Companies work with hundreds of thousands of datapoints per day, using mathematical formulas and algorithms to generate representative metrics. Metrics can be beneficial by condensing data in a more readable and digestible way and provide transparency to inform future decisions, however, like all technologies, there are drawbacks. Metrics based on skewed data can improperly reflect the composition of datapoints (Atler et. al., 2016). People's biases are reflected in the data used to create the machine learning (ML) models and artificial intelligence (AI) algorithms that generate metrics, which when used to make decisions in the real world can unfairly affect certain identity groups (Coté et al., 2021). The purpose of this paper is to analyze a potential harmful effect of metrics in a societal context, examining how using algorithms to analyze people's credit score and loan eligibility to speed up banks' decision processes can marginalize certain minority identity groups. I will use the handoff model to determine the groups involved in the creation and use of a machine learning model and to understand how bias can be introduced and propogated through the process. Additionally, I will use actor-network theory (ANT) to analyze the relationships between involved groups to discover the social forces that drive behaviors between these groups and how those relationships can be changed to prevent biased algorithms from being used.

Background

Metrics produced by machine learning and data algorithms allow companies to automate processes, meaning things that were once done manually can be done by a computer. This automatio reduces bottlenecks, makes workers more efficient and increases revenue (Atler et al., 2016). In the banking industry, companies have started using machine learning models to generate metrics which can determine who should and should not qualify for certain loans and

what percent interest to allocate for an approved loan based on their credit history and financial data. These algorithms allow bankers to make faster decisions, reduce manual labor, add consistency, reduce human error, and dynamically improve and update loan application reviews (Dakin, 2022).

However, with the adoption of models to achieve faster and more accurate decisions may come the tradeoff of fair decisions. Machine learning models bias arises when the result of a model is systematically prejudiced due to assumptions made in the process (Pratt, 2020). In a study by Coté et al., biases can appear in the decision-making process from the underlying data while creating the model and/or in the output when deploying the model (2021). Their research paper argues that an algorithm's bias is discriminatory if the differences between its result and the expected outcome given the context in which its employed are due to protected attributes, which are encoded in legislation and include attributes like sex, gender, and ethnicity, or proxy attributes, which are closely correlated with protected attributes (Coté et al., 2021). When such biases are present in a model deployed in a real world, it can perpetuate systematic racism, sexism, homophobia and more. Though banks are prohibited from considering protected attributes like race and ethnicity in making decisions under the Equal Credit Opportunity Act (What protections do I have against credit discrimination?, n.d.), the use of proxy attributes like length of credit history, neighborhood of residence and job income bracket in training models means there may be increases the likelihood of minority loan applications being rejected (Andrews, 2021). A study conducted in 2021 found that nationally, people of color were 40-80% more likely to be rejected for a mortage application than their white counterpart despite having the same credit score, with the disparity being as high as 250% in some cities (Martinez &

Kirchner). This in turn can aggravate credit inequality, causing the models to become even more biased and further marginalize traditionally lower income identity groups (Andrews, 2021).

The results found from this research can point out one of the drawbacks of using metrics without vetting the process with which they were calculated. It is important to analyze how biases in society can be displayed in data, which is used to create metrics that inform decisions that impact society, creating a self-perpetuating cycle of marginalization.

STS Frameworks

I will initially use the handoff model, which is used to describe how data and information flows through a system. The handoff model aims to show how different groups in a system interact with an artifact before it reaches the end user (Baritaud & Carlson, 2009). Within the loan approval process, there are a variety of groups that a machine learning model flows through before reaching the end user. Using the handoff model, I can determine the different groups involved in the end to end process of a loan approval algorithm being created and used. The groups identified in the loan approval algorithm pipeline can be used to create an actor-network theory to identify the other relationships that exist between groups. Handoff model can also be used to determine how bias flows down the loan approval pipeline and can inform some relationships in ANT that could be strengthened to reduce bias propogation.

Actor-network theory is a methodological approach to examining social relationships in a system. In ANT, actors are a human or nonhuman entities that influence the activity in a techno-social network, which is comprised of technological artifacts, the people that interact with them and the people's relationship with the technology (Crawford, 2020). Analyzing the relationships between the actors in a network can provide insights into the social situation that surrounds the system (Callon, 1984). ANT's primary purpose is to use evidence to describe, rather than

explain, social activity which can then be used to understand the social forces in a network (Crawford, 2020). With ANT, I can determine the social influence that different groups have on each other's behaviors and how that can prevent the filtering out of bias in the model creation and approval process.

By using handoff model, I can understand how bias arrises and is propogated in the loan approval algorithm process, and by using ANT, I can understand the social forces in this system and how they allow for bias to be introduced and maintained. With this, I can develop a more structured and transparent approach to creating, developing and implementing models with minimal bias.

Analysis

Handoff Model

Using the handoff model, I will first discern the groups involved in creating, approving and using a loan approval algorithm, as shown in Figure 1. First, data collectors gather and manipulate previous bank user loan approval data, which is then given to model designers, who define the scope of the model and its expected output. Model developers use that design to create a model, which is assessed by a testing team to ensure it meets the requirements outlined by the designers (Karani, 2023). If the model achieves the established goals, it is handed to regulatory bodies within banks to ensure it does not violate any laws, such as the Equal Credit Opportunity Act. Once approved, the model is integrated into the bank's system for bankers to use when they receive new loan applications. Whether loan applicants know it or not, their loan approval is

often determined by a computer algorithm.

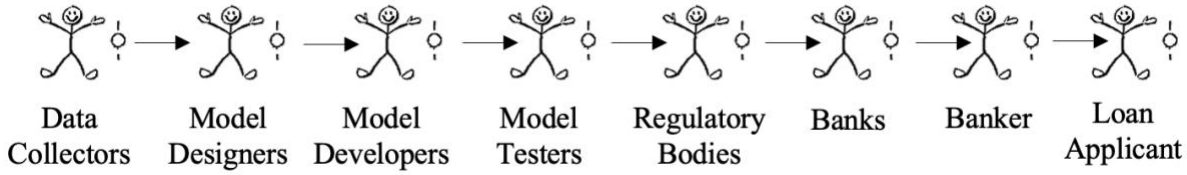

Figure 1. Loan Algorithm Handoff Model. This figure outlines the groups involved in the start to end process of developing a loan approval algorithm (Chawla, 2023).

As the algorithm moves through this development and release pipeline, bias can be introduced or propogated. For example, a commonly used technique of data cleaning that data collectors employ is removing rows with invalid or missing data. Missing/incorrect data tends to be more prevalent with minority demographics as they may not have a long credit history and may have limited access to banking systems to verify the correct input of information (Andrews 2021). If data collectors "clean" the data using this technique, minority demographic groups may have very little data and the few datapoints that do exist may skew the result of the model designed and created by the designers and developers respectively (Pratt, 2020). With the current state of testing and regulation, there are no checks in place to prevent the biased model from being propogated down to the end user. This is especially problematic as often times the metrics produced by these models when in use are converted to datapoints to feed back into the model development and training process, which can further bias the model (Carew, 2023).

Actor-Network Theory

What handoff model fails to showcase is how the different groups involved in model creation and use influence each other in a social context. While there is very little direct interaction with groups that are not adjacent to each other in the model, social forces from each group act on each other to influence behavior. Using an actor-network theory map, I can identify the other relationships that exist between the groups in the system. I can supplement

relationships between actors in ANT with the development process identified in the handoff model to implement a system of checks to ensure that if bias is introduced in the model, it does not get propogated down to the end user.

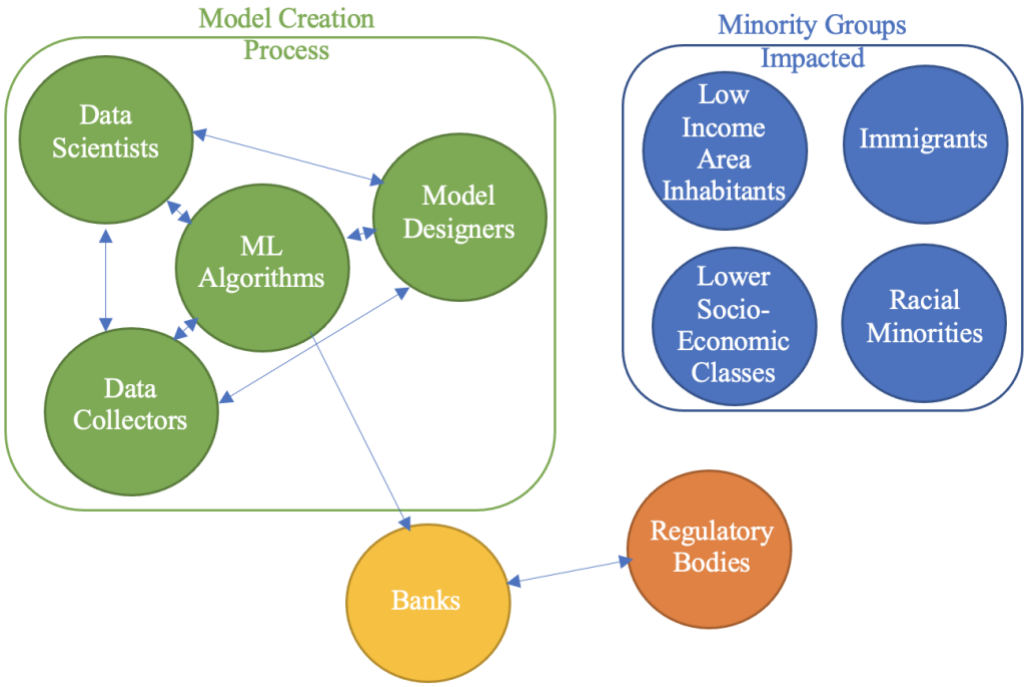

Figure 2. Loan Metrics Actor-Network. This figure outlines the different groups involved in a loan approval ML model and links between them (Chawla, 2023).

The actor-network theory map, seen in Figure 2, can be used to analyze how disconnects in information and collaboration can bring rise to bias (Callon, 1984). Data scientists, model designers and data collectors all work together to create ML algorithms, but their workings are a black box to the other actors in the network. Banks use the end-product algorithm without much knowledge about the creation and testing processes (Rudin & Radin, 2019). This is especially problematic because those that are involved in the model creation process do highly technical work and often do not consider the social implications of the models they are creating (Martin & Moore, 2020). Without a deeper knowledge of how these models work, banks cannot determine whether the output model from the model creation process is biased.

Additionally, while bank regulatory bodies like the FDIC work to ensure that customers and clients are treated fairly and honestly, the only regulation relating to machine learning models enforced is that protected attributes like race, gender and ethnicity cannot be used in making decisions (FDIC, 2022). This is not sufficient in preventing bias from being introduced to models as proxy attributes that are closely correlated with protected attributes still can be used and can lead to models outputting discriminatory metrics (Coté et al., 2021).

The end user minority groups impacted by these models are not consulted by any of the other actors in this network yet are the most affected by these biased models. Without banks, regulators and data engineers understanding their end user groups and how such models impact them, it is difficult to ensure that metrics generated by loan approval models are equitable across demographic groups (Brown, 2022).

Discussion

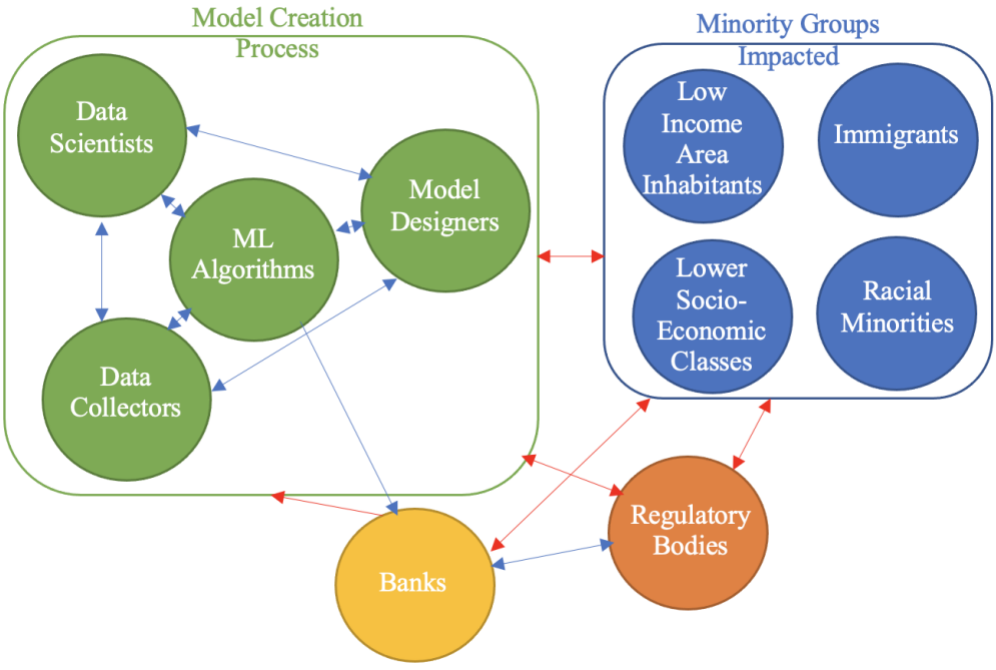

Figure 3. Loan Metrics Actor-Network Supplementary Relationships. This figure outlines the relationships that should be bolstered in the loan metric system to reduce biased models (Chawla, 2023).

To mitigate the propagation of bias in the pipeline established in the handoff model, I believe that the existing interactions outlined in blue in Figure 3 should be supplemented with those in red. In involving the minority end users in the model creation process, the development of biased models can be prevented. Having an understanding of how historical data may perpetuate historical prejudices towards minority groups may alter how data collectors find and manipulate data. Additionally, it may also prevent model designers and creators from using proxy attributes in the development of the model, further mitigating the presence of bias. For example, currently, models use the applicants assets beyond the down payment as a factor in determining whether the loan should be approved. While it is true that those with more assets are more likely to pay back the loan, with the understanding that minority racial groups often have less assets than their white counterparts since they don't benefit from intergenerational wealth as much and engage with the stock market less, analysts can run correlation analyses to determine if some methods of calculating assets can produce proxy attributes. If so, model developers can leave it out of the model creation process or give it less weight in the model to prevent the use of a feature associated with the protected attribute of race (Martinez & Kirchner, 2021).

Increasing information flow between banks, the model creation and development groups, and the end users could prevent the use of already biased models and ensure that decisions made by algorithms do not disproportionately impact minority groups. Explainable machine learning is a new branch of data science that aims to tear down the black box that surrounds algorithms and models. The goal is to create models that can provide evidence from the decision making algorithm and underlying data to explain how a model arrives at its output metric (Onose, 2023). Improving training and use of explainable machine learning techniques would allow banks to

better understand the workings of the model creation process and could add another check to ensure that created and tested biased models do not reach the end user.

Regulatory bodies are currently the most removed from the system but may have the most power to dictate the future scope and use cases of machine learning models in banking. There have been some increases in regulatory body involvement with ML/AI in the financial space, which have lead to more financial firms increasing their transparency in their algorithm creation and use (Hutto-Schultz, 2021). Extending this to banking could lead to more effective regulations on the use of loan approval ML models and metrics.

Conclusion

While metrics produced by machine learning models have the potential to optimize company workstreams, increase efficiency and reduce human error, they also have the potential to perpetuate the biases in the models' underlying data. In the banking industry, loan approval algorithms have been found to disproportionately deny loans to applicants of certain minority and ethnic groups due to limited historical data about their affiliated information like lower income neighborhoods and short credit histories and using attributes closely affiliated with protected attributes like race and gender to make decisions in the model. Using handoff model, I determined the groups involved in creating and approving a loan approval model, and identified where bias can arise and propogate. I then used actor-network theory to identify how relationships between these different groups can influence behavior and potentially mitigate bias. Understanding the social forces within this system via ANT can allow for better accountability between the different groups involved in the system and can result in the prevention of the use of biased models.

REFERENCES

Andrews, L. E. (2021, August 6). How flawed data aggravates inequality in credit. *Stanford University Human-Centered Artificial Intelligence.* https://hai.stanford.edu/news/how-flawed-data-aggravates-inequality-credit

Atler, A., Sachdev, S., Wilson, H. J. (2016, May 3). How companies are using machine learning to get faster and more efficient. *Harvard Business Review.* https://hbr.org/2016/05/how-companies-are-using-machine-learning-to-get-faster-and-more-efficient

Baritaud, C., Carlson, B. (2009). STS Handoff Model. *Class handout* (Unpublished). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Bijker, W. E., & Pinch, T. J. (1984). The social construction of facts and artifacts. *Social Studies of Science*, 14, 399–441. https://doi.org/10.1177/030631284014003004

Brown, S. (2022, January 26). For successful machine learning tools, talk with end users. *MIT Sloan.* https://mitsloan.mit.edu/ideas-made-to-matter/successful-machine-learning-tools-talk-end-users

Callon, M. (1984). Some elements of a sociology of translation: Domestication of the scallops and the fishermen of St Brieuc Bay. *The Sociological Review*, 32, 196-233. https://doi.org/10.1111/j.1467-954X.1984.tb00113.x

Carew, J. M. (2023, February). Reinforcement learning. *Tech Target.* https://www.techtarget.com/searchenterpriseai/definition/reinforcement-learning

Chawla, K. (2023). *Loan Algorithm Handoff Model.* [Figure 1]. *STS Research Paper* (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Chawla, K. (2023). *Loan Metrics Actor-Network.* [Figure 2]. *STS Research Paper* (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Chawla, K. (2023). *Loan Metrics Actor-Network Supplementary* Relationships. [Figure 3]. *STS Research Paper* (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Coté, M., Criado, N., Ferrer, X., Nuenen, T. van, & Such, J. M. (2021, June 2). Bias and discrimination in AI: A cross-disciplinary perspective. *IEEE Technology and Society Magazine*, 40(2), 72–80. https://doi.org/10.1109/MTS.2021.3056293

Crawford, T. H. (2020, September 28). Actor-Network theory. *Oxford Research Encyclopedia of Literature.* https://doi.org/10.1093/acrefore/9780190201098.013.965

Dakin, R. (2022, January 15). How banks are using machine learning. *Bits in Glass.* https://bitsinglass.com/how-banks-are-using-machine-learning/

FDIC. (2022, October 11). Other Regulators and Organizations. *FDIC.* https://www.fdic.gov/resources/consumers/other-regulators.html

Hutto-Schultz, A. (2021, April 5). Federal regulators' artificial intelligence initiative is a promising development for financial industry. *Davis Wright Tremaine LLP.* https://www.dwt.com/blogs/privacy--security-law-blog/2021/04/federal-financial-institutions-ai-rfi

Karani, D. (2023, January 26). Roles in ML team and how they collaborate with each other. *Neptune.ai.* https://neptune.ai/blog/roles-in-ml-team-and-how-they-collaborate

Lynch, S. (2015, June 26). The benefits of having the right metrics (key performance indicators). *Linkedin.* https://www.linkedin.com/pulse/benefits-having-right-kpis-key-performance-indicators-stephen-lynch/

Martin, D., & Moore, A. (2020, October 28). AI engineers need to think beyond engineering. *Harvard Business Review.* https://hbr.org/2020/10/ai-engineers-need-to-think-beyond-engineering

Martinez, E., & Kirchner, L. (2021, August 25). The secret bias hidden in mortgage-approval algorithms. *The Markup.* https://themarkup.org/denied/2021/08/25/the-secret-bias-hidden-in-mortgage-approval-algorithms

Onose, E., (2023, April 19). Explainability and auditability in ML: Definitions, techniques, and tools. *Neptune AI.* https://neptune.ai/blog/explainability-auditability-ml-definitions-techniques-tools

Pratt, M. K. (2020, June). Machine learning bias (AI bias). *Tech Target.* https://www.techtarget.com/searchenterpriseai/definition/machine-learning-bias-algorithm-bias-or-AI-bias

Rudin, C. & Radin, J. (2019, November 22). Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. *Harvard Data Science Review.* https://hdsr.mitpress.mit.edu/pub/f9kuryi8/release/8

What protections do I have against credit discrimination? (n.d.) *Consumer Financial Protection Bureau*. https://www.consumerfinance.gov/fair-lending/