Investigating Algorithmic Detection of Welfare Fraud

Dylan Van Bramer, Madeline Demers, Tanvi Namjoshi, Ella White

Research Domain

In this project, we plan to analyze the use of algorithmic decision-making systems in welfare administration. Algorithms are used globally to make decisions about welfare from allocation to fraud detection. In our analysis, we focus specifically on the use of risk-score algorithms that attempt to predict which beneficiaries of welfare payments should be investigated for fraud. We found instances of such algorithms being used around the world. Our literature review will focus on three specific case studies: the Netherlands, Michigan, and France. The articles in our literature review all refer to discrimination based on a large list of protected features. However, some common sensitive features emerged across all cases including age, ethnicity, single-parent status, and disability status. In our analysis, we plan to focus on age, although our audit will seek to confirm some of the results regarding ethnicity and parental status. In the literature, the latter two features were often not explicitly in the training data but rather evident through proxies.

Millions of people across the globe rely on social safety nets and welfare to provide for themselves and their families. Socially, we should care about who is targeted for audits because they can be extremely disruptive to the lives of the family that is targeted. Audits are invasive and potentially traumatizing for families, requiring many documents to be turned over, neighbors to be interviewed, and statements to be scrutinized. States should also care about the fairness and accuracy of their systems. When the algorithms used to trigger investigations are wrong it wastes state resources that could be put to better use.

Literature Review

Rotterdam

In an article titled, “This Algorithm Could Ruin Your Life”, WIRED reporters worked in collaboration with the informational magazine Lighthouse Reports, the Pulitzer Center’s AI Accountability Network, and the Eyebeam Center for the Future of Journalism. This article from June of 2023 is an investigative journalism piece written by Gabriel Geiger, Evaline Schot, Matt Burgess, and others. The article provides a detailed examination of the implementation and consequences of a machine learning algorithm used to predict welfare fraud in the city of Rotterdam, located in the country of the Netherlands. It discusses how the algorithm assigns risk scores to welfare recipients based on various factors, including demographic information and personal history. Through interviews and personal stories, the article illustrates the negative impact of algorithmic decision-making on individuals, particularly vulnerable populations such as single mothers and those with health concerns. It highlights concerns about transparency, fairness, and potential biases in the algorithm's operation, as well as the ethical implications of relying on such technology in social welfare systems [1]. The assumptions presented in the paper seem valid based on the information provided. The main contributions of the paper are shedding light on the use of algorithms in social welfare systems, highlighting the potential biases and ethical issues associated with their implementation, and raising awareness about the impact on vulnerable individuals. The paper is well-written, providing clear explanations of complex concepts and presenting the information in a structured and engaging way. Furthermore, the inclusion of personal stories adds depth and relatability to the topic. The comprehensive dive into this case study is a particularly interesting example that is elevated by the journalistic nature of the piece and the inclusion of personal experiences from different perspectives of those affected by the algorithm. While the article reveals the negative aspects of this algorithm, noting in the title that it could ruin the lives of individuals, the piece could benefit from further exploration into the communication between the developers of such algorithms and those using them to assess cases of potential fraud. The article appears to narrate a story that targets the "black box" nature of the algorithm to be problematic, using 315 features to determine a risk score, but not explaining exactly how. This question of the divide between humanity and algorithms in the case of decision-making is a key question that relates to the overall inquiry into the implications of using algorithms to provide information that contributes to decision-making about people's lives. The factors explored in this analysis relate to the other literature review pieces since they all explore the topic of social welfare and the use of algorithms, identifying a clear need to study the connection between these two topics.

Rotterdam Data Analysis

The article titled "Suspicion Machine Methodology" is an investigative journal article as a collaboration with journalists from WIRED, the Dutch Public Broadcasting System VPRO, the Pulitzer Center, the nonprofit Follow the Money, the magazine Vers Beton, and Open Rotterdam. This piece was written by Justin-Casimir Braun, Eva Constantaras, Htet Aung, Gabriel Geiger, Dhruv Mehrotra, and Daniel Howden. This piece goes further in-depth with the data analysis of the algorithm in the WIRED article above. The algorithm in question was a gradient-boosting machine model deployed by Rotterdam to identify welfare recipients who were cheating the welfare system. Based on 315 inputs, this algorithm selects approximately one thousand "riskiest" recipients for investigation. Rotterdam's system is one of many automated systems deployed by governments across the world. Since code is often proprietary and there are often strong data privacy concerns, auditing government use of algorithms like these is difficult. With access to the model file, training data, and code for the system, this paper provides a uniquely close look at such a system. Lighthouse Reports found that the Rotterdam model was fifty percent more accurate at predicting fraud than selecting people at random. In terms of fairness metrics, Lighthouse focused on statistical parity and controlled statistical parity. It was also found that the algorithm was harsher on those who are "parents, young people, women, people with roommates, people who do not have enough money and people with substance abuse issues" [2]. In addition, those who were not proficient in Dutch were 2.22 times more likely to be flagged than those proficient. With language as a proxy for ethnicity and race, such data is concerning. Similarly, age was the most important factor in determining a risk score, a concerning pattern since age is a factor that is beyond one's control.

One major limitation of this paper was that no information was provided about actual fraud rates of certain subgroups, so it was difficult to consider fairness metrics like predictive equality. Further, around half of those in the dataset had committed fraud whereas in the real world that is closer to 21%. The co-authors were transparent about these limitations as well as how calculations were considered for the data analysis. It would be beneficial to understand how this model fits in with other past models. Notably, we cannot assume this data is fully representable, for it selects only those who live in Rotterdam, which is a self-selecting group. To mediate the potential for these effects to agglomerate downstream, the investigators utilized conditional parity. For instance, the effect of being a parent on ratings was tested based on an experiment that created two copies of the training dataset, one with everyone meant to be a parent and one with everyone not meant to be a parent. From this, it was concluded that those who were parents were overrepresented by a factor of 1.09 in the high-risk group.

These skewed findings provide compelling grounds to further investigate these algorithms and take a critical eye on deploying algorithms in these contexts where often the agencies who deploy the model do not understand the technology itself.

France

In the investigative journalism piece by Le Monde, the authors explore the algorithm used by the Caisse Nationale des Allocations Familiales, France's family-oriented welfare institution, to detect instances where the risk of fraud is high. Rather than attempting to detect when welfare fraud has occurred, the data mining algorithm assigns each beneficiary a risk score that determines whether an agent will be sent to audit them. In the article, the authors present their findings from analyzing the code used by the CNAF. They found that the household risk score (between 0 and 1) is based on a logistic regression that associates around 40 features with individual risk coefficients. These 40 features include protected criteria such as age, disability, and marital status. Indirectly, the algorithm also penalizes single parents by checking whether a household has more than 15 months of combined employment in a given calendar year, a goal that is impossible for single individuals. The authors argue that the algorithm's simplicity results in threshold effects: having a spouse turn 60 can lead to a sudden jump in your household risk score, despite there being no other changes. Similar to the other literature on welfare fraud algorithms, this piece focuses on the analysis of a specific implementation in a specific location. While the fragmented nature of the literature may make it hard to provide overarching guidelines and findings, it is important as each institution uses different algorithms in its welfare system.

However, one big gap in the analysis is that, unlike the Rotterdam piece, the authors did not test the algorithm's performance on a large data set, instead focusing on a couple of individual stories to demonstrate the harmful nature of the algorithm. While it is useful to see specific cases analyzed, it would be nice to have statistics such as the percentage of single parents classified as high-risk vs two-person households. These statistics would allow us to understand whether the algorithm fulfills common fairness metrics such as statistical parity. Another gap emerges because we do not know the true outcomes of whether audited/unaudited households engaged in fraud, which means it is not possible to analyze false positive and negative rates. Despite this, one thing the authors do well is explain the impact of an audit. They explain that although this tool is not a final decision-maker, and instead triggers a human audit, the impact on families is still serious as audits are invasive procedures that leave family members subject to intense scrutiny. Finally, it is important to note that the algorithm the authors analyzed in this piece is not the most current implementation in use, and instead is from a couple of years ago. Thus we cannot assume that any discriminatory behavior found is currently used in practice. When the CNAF sent the current algorithm, they redacted the feature names from the code, thus making it impossible to analyze. This highlights a key hurdle in investigating algorithms used by governments: it is hard to gain access to information about the algorithms because agencies have security and privacy concerns.

Michigan

In the article *"Computer Says No!": The Impact of Automation on the Discretionary Power of Public Officers*, Dr. Doaa A. Elyounes focuses on policy approaches to adopting automated algorithms for welfare distribution and monitoring. Indeed, Dr. Elyounes focuses on analyzing the legal and moral structures within which algorithms are developed and deployed in this article (published by the Vanderbilt School of Law), using case studies from the Netherlands and Michigan.

The Michigan case study, in which the Michigan Unemployment Insurance Agency (UIA) deployed a new system for overseeing welfare decisions, is used, essentially, as a crash course in “what not to do.” The system deployed, the Michigan Integrated Data Automated System (MiDAS), linked data from different governmental departments (employment records, medical records, tax documents) to collect as much data as possible, and nearly autonomously handle welfare fraud detection. In the process of deployment, MiDAS replaced four hundred employees who otherwise would have reviewed documents by hand, and held interviews with individuals involved in a specific case.
Alternatively, MiDAS attempted to communicate with the individual once their profile was initially flagged. If their response to this contact was “deemed insufficient,” the algorithm automatically flags the case as fraudulent, giving the system itself discretion to cut the individual’s benefits and seize tax refunds. While this had the potential to reduce costs, as well as human bias in the process of fraud detection, the algorithm failed horribly. “Approximately 93 percent of cases were wrongly flagged,” with most of those being false positives. Because the algorithm had near-total control over handling cases, these cases weren’t just flagged – real people had their unemployment welfare cut off by an algorithm. If, on the contrary, MiDAS had served as a data analysis tool that provided suggestions to human case workers, it is possible that the individuals wrongly accused of fraud would not have faced real-life economic burden.

The main contribution of this piece is the declaration (with significant supporting evidence) that “meaningful discretion” by humans must be maintained, even when using largely automated systems for welfare decisions (and other high-stakes predictions “in the wild”). In the words of the article, these algorithms should be developed with the goal of being “decision-aiding” models, rather than entirely autonomous “decision-making” ones (like MiDAS). One of the classifications under which Dr. Elyounes analyzes policy decisions is as either “street-level” or “system-level,” where the former describes person-to-person interactions like those of individual case workers, and the latter more bureaucratic decisions like those of automated algorithms. Of course, the development of nearly autonomous algorithmic systems like MiDAS skew this balance heavily toward “system-level” policies, and risk the wellbeing of people impacted by these choices.

Compared with all of the other data-heavy pieces in this literature review (and within the field of algorithmic fairness, at large), this policy-focused piece provided interesting perspectives about algorithmic development and deployment. Despite being focused on two particular case studies, the focus on high-level policy around algorithm development and deployment made the article feel more “future oriented” and generalizable, than a data analysis of a certain predictor (though, those of course are necessary, as well). While it is unclear if this is necessarily a limitation, it is interesting that Dr. Elyounes has a very palpable opinion in the piece. Using words like “draconian” to describe the algorithms and decisions made, the author suggests moral correctness in human discretion. While many *do* agree that the adoption of entirely automated systems, especially in high-risk use cases, is a deeply moral issue, the tone of the article reinforces this heavily. However, the author presents statistical benefits of system-level decisions, as well, reinstating the objectivity of the piece.

Overall Analysis

Welfare administration is a key service that many families depend on. Having an effective method to weed out fraud is essential to ensure the system works smoothly. An algorithm has the potential to identify such instances of fraud at a larger scale, faster and more efficiently than a human team could. However, using an algorithm in such a sensitive setting, and often training it on sensitive features, introduces high risk for bias in this context where a mistake has high cost. In the case of France and Rotterdam, we see that households that undergo more financial strain, such as single family households, are the very households that get penalized by these algorithms. The case of Michigan emphasizes that these algorithms, left without any human intervention, have the potential to cause more harm than good, penalizing innocent people for fraud they didn’t commit.

Dataset Exploration

The data sources we plan to use for the project to explore the use of algorithms as it pertains to the detection of social welfare fraud is from the Surveillance Newsroom at Lighthouse Reports GitHub repository, which houses the raw data used in the project that investigated the use of algorithms by the city of Rotterdam to predict social welfare fraud in the population. The raw data is synthetic data generated for replication, due to the fact that the original data cannot be used because of GDPR concerns. The contributors to this repository are Justin-Casimir Braun, Htet Aung, Gabriel Geiger, and Eva Constantares. We believe that this data is a trustworthy source because there is extensive evidence that the data was sourced ethically and the article from Lighthouse Reports details the legal process and cases that allowed this data to be used in an effort to contribute to openness and transparency.

Link: https://github.com/Lighthouse-Reports/suspicion_machine/tree/main

Using this synthetic data, we aim broadly to answer two questions: first, are the social welfare fraud algorithms currently deployed *fair*; and, more broadly, can a *fair* automated algorithm for welfare detection even be created?

To dive more deeply into the first question, we will conduct extensive data analysis on the classifier deployed in Rotterdam. While Lighthouse Reports has already conducted algorithmic analysis including statistical parity and conditional statistical parity, we aim both to confirm their results externally, as well as to introduce other measures of fairness (calibration, equalized odds, and more). Furthermore, we will conduct each of these analyses across more subgroups and protected features, particularly focusing on intersectionality (looking at age AND gender concurrently, for example, and seeing if biases can be thought of "stacking" mathematically in an additive or multiplicative fashion).

To explore the second question more thoroughly, we will first attempt to train our own predictors using different machine learning algorithms than used in the current model. We will train and test a variety of these, largely focusing on interpretable, classic ML methods like RandomForests and SVMs. Once we attain the results from our own model training and testing (and analyze our models using the same criteria we applied to the Rotterdam algorithm), we will then reflect more qualitatively on the viability of automated predictors in a welfare setting. Referencing the literature above, as well as other foundational fairness papers, we will explore alternatives to automation, and emphasize the importance of "humans in the loop" during algorithmic development and deployment.

The dataset is large enough to do these analyses and it contains specific variables for the sensitive features. The Lighthouse Reports article does an analysis of the data and the model and provides helpful resources on how to experiment in a new way for those seeking to further understand and investigate the case. The research that they did is something we aim to replicate. We then expect to dive deeper into the model in order to explore other aspects of the data, specifically instances of intersectionality in relation to demographic details.

Some of the limitations of the data set are notable mentions to this project. Firstly, the column names are in Dutch, which we plan to translate using a resource such as Google Translate to use the English translations in our analyses. Secondly, the data is not perfectly representative of the actual fraud rate in the city of Rotterdam, which is around 21%. The data in the sample has a fraud rate of about 50%, which is much higher than the fraud rate in reality. This is because only a partial amount of the training data from the city was accidentally sent to the investigative team at Lighthouse Reports, which they were able to legally use after extensive legal proceedings and a landmark decision allowing them access to the data that was leaked.

Contribution Notes

In this phase submission, the group contributions were distributed as follows: Madeline completed the investigative journalism piece "This algorithm could ruin your life" from *WIRED UK.* Ella completed the Rotterdam Data Analysis section of the literature review. Tanvi completed the literature review for France's CNAF risk score system. Dylan completed the literature review regarding the Michigan Integrated Data Automated System (MiDAS) using the Vanderbilt law article. The remainder of this phase including the general research stage, research domain section, and dataset exploration section was worked on collaboratively by all group members, while in meetings together.

Bibliography

[1] Matt Burgess, Gabriel Geiger, and Evaline Schot. 2023. This algorithm could ruin your life. (June 2023). Retrieved February 10, 2024 from https://www.wired.co.uk/article/welfare-algorithms-discrimination

[2] Justin-Casimir Braun, Eva Constantaras, Htet Aung, Gabriel Geiger, Dhruv Mehrotra, and Daniel Howden. 2023. Suspicion Machine Methodology. Lighthouse Reports. Retrieved February 19, 2024 from https://www.lighthousereports.com/methodology/suspicion-machine/

[3] Manon Romain, Adrien Sénécat, Elsa Delmas, Thomas Steffen, Léa Girardot, and Lighthouse Reports. 2023. Is Data Neutral? How an Algorithm Decides Which French Households to Audit for Welfare Fraud. (December 2023). Retrieved February 18, 2024 from https://www.lemonde.fr/en/les-decodeurs/visuel/2023/12/05/how-an-algorithm-decides-which-french-households-to-audit-for-benefit-fraud_6313254_8.html

[4] Doaa A. Elyounes. 2021. "Computer Says No!": The Impact of Automation on the Discretionary Power of Public Officers. *Vanderbilt Journal of Entertainment and Technology Law* 23, 3 (Spring 2021). Retrieved from https://scholarship.law.vanderbilt.edu/jetlaw/vol23/iss3/1