Algorithm Audit

Problem statement

Risk Profiling for Social Welfare Re-examination

Algoprudence
AA:2023:02:P

May 2023

Abstract

This problem statement describes several ethical issues relating to manual and algorithmic profiling methods used by Dutch municipalities to sample citizens in the context of social welfare benefits. Information about these methods has been collected and analyzed independently by Algorithm Audit. This document has been strengthened by feedback received from experts with various professional backgrounds. As a starting point, data on the profiling algorithm was retrieved through a freedom of information request by journalists, e.g., source code, model documentation, Data Privacy Impact Assessment and aggregated performance statistics. Over time, additional legal, statistical and context-related information was added to the problem statement, some of which was supplied by the Municipality of Rotterdam upon the request of Algorithm Audit. The profiling algorithm here under review is no longer used by the Municipality of Rotterdam. Yet, the underlying ethical concerns remain relevant in the wider debate on risk profiling and its applications throughout society. Based on this document, an independent audit commission gives advice on three identified ethical issues. This advice can be found in corresponding Audit Commission Advice Document (AA:2023:02:A).

Table of contents

Problem statement – Risk Profiling for Social Welfare Re-examination

This document describes ethical concerns regarding risk-based profiling methods to select social welfare recipients for re-examination. Real-life sampling methods as used by the Municipality of Rotterdam are analyzed in this case study.

Introduction

Not all social welfare is granted lawfully. Unawareness of responsibilities, administrative mistakes, culpable negligence, and fraud result in unlawful payments. Regardless of the underlying reason, Dutch municipalities have a statutory duty to detect and reclaim unduly granted social welfare payments. Dutch municipalities therefore perform regular re- examinations of welfare recipients to review whether social welfare claims are duly granted. These re-examinations take place in the form of interviews and human checks of provided documents. The purpose of these re-examinations is not only to exclusively detect fraud, but also to detect unintentional administrative mistakes and other errors that result in unduly granted welfare payments, e.g., updating outdated information.

If selected for re-examination, recipients are summoned to an interview with a civil servant, to check whether the information the municipality holds about the recipient is up to date. The municipality perceives a re-examination interview as a natural moment of contact with citizens without *a priori* suspicion. Nonetheless, interviewees regard the period towards and the re-examination interview itself as invasive, time-consuming and stressful, due to among others the administrative burden, e.g., handing over identity documents, diplomas, or providing an overview of personal spending to civil servants. In Box 1, a testimony of a recipient invited for a re-examination interview is stated. Recipients are expected to plan

About Algorithm Audit

Algorithm Audit is a European knowledge platform for AI bias testing and normative AI standards. Its activities are three-pronged:

Audit commissions

Advising on ethical issues emerging in concrete algorithmic practices through deliberation, resulting in *algoprudence* (see Box 2)

Technical tooling

Implementing and testing technical tools to detect and mitgate bias in data and algorithms

Knowledge platform

Bringing together knowledge and expertise to ignite the collective learning process for reponsible algorithms

an appointment with a civil servant within two weeks after having received an invitation letter.

As not all recipients can be re-examined, sampling methods are used to select recipients for re-examination. To allocate labor resources efficiently and to help recipients in an early stage with unduly granted payments, municipalities operate risk-based sampling methods to select recipients for re-examination. The goal of risk-based sampling is to invite those recipients for an interview, for which there is a greater likelihood that social welfare is unduly granted. Various risk-based sampling methods exist. For instance, civil servants working as a personal client manager or income counselor can short-list recipients for re-examination if interaction arouses suspicion. Citizens can also contact the municipality if they have suspicions about fraudulent behavior of others, which is validated by a civil servant[1]. This selection method is called *event-driven sampling*.

When no such event occurs, various other risk-based sampling methods are used. For instance, recipients are selected on the basis of risk profiles. These risk profiles are generated according to a set of pre-defined selection criteria, which are supposed to reflect the characteristics of recipients with a higher risk of receiving undue benefits on average. The criteria for such risk profiles are derived in multiple ways. Based on professional experience, subject matter expertise (SME) of civil servants is used to manually define criteria for risk profiles, e.g., men living alone[2]. This method is called *SME profiling*. A SME profile typically consists of 1-3 selection criteria that are manually defined and, most often, are annually changed to avoid overemphasis on a specific group. More details on the work process of SME sampling in Rotterdam can be found in Appendix C – Variable selection methods.

A different method to build risk profiles is *algorithmic profiling*. In this approach, a large amount of historical data on characteristics of recipients and the outcome of re-examination interviews is used to generate risk profiles in an automated manner. The model is trained on the target variable 'unduly granted welfare payments', which is established through conducting re-examination (more details in section Legal background of social welfare re-examinations). Once established, either through SME or algorithmic profiling,

[1] Hotline of the Municipailty of Rotterdam to report suspected behaviour https://www.rotterdam.nl/loket/fraude-uitkering-doorgeven/

[2] Colored Technology, Rotterdam Court of Auditors 2021 https://rekenkamer.rotterdam.nl/onderzoeken/algoritmes/

Scale of unduly granted social welfare

In 2017-2021, for 38% of the approximately 22.000 re-examined social welfare recipients in Rotterdam action was required to make the payments duly granted, through mutation or termination of payments, or through administrational adjustments. In 2019, the Municipality of Amsterdam and the Municipality of The Hague observed in total respectively €4.9M and €5.5M of unduly granted payments. For Rotterdam these numbers are not available. More aggregation statistics are provided in Appendix D – Sampling performance metrics.

the risk profiles are subsequently applied as a selection filter to sample recipients. Lastly, through *random sampling* recipients are selected in a random manner, without any selection criteria or risk profiles involved. These various sampling methods are used by Dutch municipalities to select recipients of social welfare for re-examination. A representation of these sampling methods is given in Figure 1. Proportions of sampling methods vary per year and per municipality. More details on sampling proportions in the Municipality of Rotterdam in 2017-2021 can be found in Appendix D – Sampling performance metrics.

Scope of case study

In this case study, ethical concerns pertaining to algorithmic sampling and SME sampling methods are examined, in regard to their variable selection and risk-profiling methods. A first ethical risk concerns biased historical data from which selection criteria for risk profiles are distilled. A second ethical risk relates to the variable selection process itself, performed either manually through SMEs or automatically through an algorithm. For reasons explained below, we focus in this case study on the second rather than on the first ethical risk.

Based upon a freedom of information request[3], a consortium of investigative journalists has recently obtained access to the datasets that were used by the Municipality of Rotterdam to train the risk-prediction algorithm. Their investigation has shown that these training

[3] FOI request VPRO Argos/Lighthouse Reports, 2017020 Privacy Impact Assessment Duly granted Social Welfare https://www.vpro.nl/dam/jcr:c87f2d6c-3f9c-4498-9a9c-f3bc5483a437/Downloads%20Model%20Rotterdam.zip

Box 1

Stakeholders consulted for this project

For this project the following stakeholders are heard:

- Individuals subjected to the algorithm
- Representatives of affected groups
- Municipality of Rotterdam
- Research journalists
- Municipal institutions (Rotterdams Court of Auditors and Ombudsman of Rotterdam)
- Legal experts and academic researchers

"Being invited for an interview is stressful. If you don't show up at the appointment, as stated in an unannounced invitation letter, your payments are automatically terminated. That I, as a born and raised Dutch citizen, was asked to prove that I am proficient in the Dutch language was humiliating.."

Social welfare recipient from Rotterdam (anonymized)

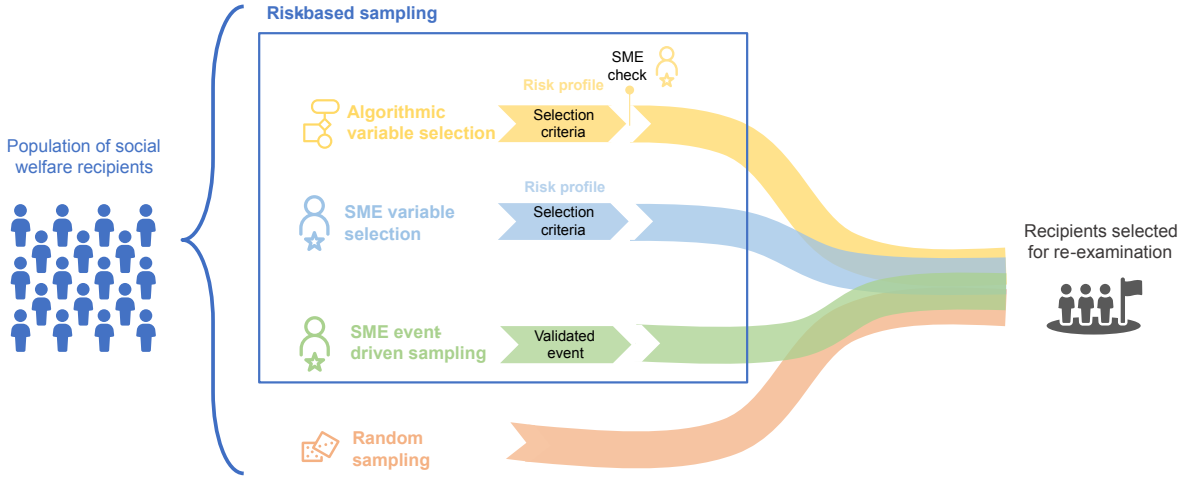

Figure 1 – Sampling methods to select social welfare recipients for re-examination

datasets are biased towards age and gender[4]. Identification of unfair data is important to prevent 'garbage in, garbage out' statistical learning. But this does not address *qualitative* concerns around risk profiles and the use of selection criteria, including SME methods that are less data- driven. Even when datasets are considered to be fair, or in case data-driven methods are not used, qualitative concerns on selecting risk profile criteria remain. These qualitative concerns are at the heart of this case study. Another reason to leave the quality of the data aside, is that it is highly specific to the data set at issue. This means it is difficult to generalize assessments of data quality to other municipalities, compared to the qualitative issues which apply to all potential use cases. Furthermore, the Municipality of Rotterdam has set up an internal ethics board, which will have a mandate and will be better positioned to perform audits on the data quality, for instance to assess biases in the historical data.

Before elaborating on the research questions of this problem statement, it is important to clarify the current status of the Municipality of Rotterdam's sampling algorithm. After publication of the report *Colored Technology* by the Rotterdam Court of Auditors in April

[4] See https://www.lighthousereports.com/suspicion-machines-methodology/

Box 2

Process of Algorithm Audit's case studies

Problem Statement

> Identify ethical issue(s)
> Collect relevant information
> Process feedback of experts

Audit commission

> Compose diverse audit commission
> Initial written feedback on Problem Statement
> Deliberative discussion on ethical issues

Algoprudence

> Compose normative advice based on gathering
> Approval of all audit commission members
> Publication of algoprudence

2021, members of the city council addressed concerns on algorithmic sampling and the process of re-examination to the responsible councilor. As a result, later in 2021 application of the algorithm was stopped[5], and social welfare recipients continued to be sampled either through event-driven sampling, SME and random sampling. Over the past years, the Municipality of Rotterdam has been developing a new algorithm to sample welfare recipients, which is supposed to meet higher standards than the previous model. The status of this newly developed model is unknown.

Ethical issues

We now move to the specific focus questions of this case study. A first focus of this case study is the issue of proxy discrimination in risk profiling. Using apparently neutral characteristics of citizens (such as level of education, ZIP-code or fluency in Dutch) as selection criteria for risk profiles, may induce discriminatory outcomes as a result of their correlation with protected characteristics, such as ethnicity. The question then becomes what characteristics, taking into account their potential proxy-character, can still legitimately be used, and which must be excluded.

A second ethical risk concerns the use of variables that are morally doubtful or problematic as such. As described below, some of the personal data used for profiling tend to be based on subjective assessments by civil servants, or appear to be personal traits extraneous to the aim of selection for re-examination. With respect to using these criteria in profiling methods, it can be questioned whether criteria are objective, proportionate and necessary with regard to the aim pursued.

A third issue concerns the difference between algorithmic profiling and SME profiling. Does SME sampling increase or decrease the ethical risks relative to algorithmic sampling? Are the ethical risks of profiling of welfare recipients primarily located in the use of

[5] See https://www.ftm.nl/artikelen/algoritme-gemeente-rotterdam

Algoprudence: Public knowledge building for ethical algorithms

Algorithm Audit does not have a mandate to issue legal rulings or official judgements. In our case studies, we give non-binding ethical advice. Ethical advice often goes beyond advice on what is required for legal compliance. Yet in the absence of legal rulings or clear standards established by a supervisory body, our independent ethical advice also serves as a preliminary signpost for organizations. Our case advice may also help elaborate official standards or support future decisions by legal bodies. In this sense, our ethical advice does have relevance for the legal domain.

machine learning techniques for variable selection – or are they located in risk-based profiling itself, which means that manual SME profiling suffers the same risks as algorithmic profiling?

The scope of this case is hence three-pronged. We establish the following ethical issues that guide this case study:

> Issue I: What characteristics of recipients can be considered as a proxy variable for protected attributes (as defined in Article 14 of the European Convention on Human Rights), and which of those variables should be excluded from profiling methods to mitigate discriminatory bias?
>
> Issue II: What characteristics are ethically undesirable to use in profiling methods, for reasons other than discriminatory bias?
>
> Issue III: Under what circumstances is it desirable to select recipients for re-examination through algorithmic sampling, rather than by SME sampling?

In the forthcoming sections, the mentioned ethical issues are elaborated on in more detail. Additionally, details on the legal basis, available selection criteria, and sampling practices are discussed.

Legal background of social welfare re-examinations

Before assessing sampling methods in detail, it is relevant to question whether the use of risk- based profiling methods for re-examination of social welfare payments is legally justified. As described by the Dutch College for Human Rights in the report *Discrimination through Risk Profiles*[6], Dutch and European courts have ruled that risk profiling techniques used for effective, efficient, or cost-saving social welfare re-examination sampling, serve a legitimate objective and can be legally justified[7]. Nonetheless, if used for risk-based sampling, profiling methods need to obey among others to the right to non-discrimination, equal treatment, data protection and privacy laws. One way in which these fundamental rights are guaranteed, is that differently sampled recipients are treated alike, i.e., civil servants who conduct re-examination interviews do not know how interviewees are sampled, e.g., through random, SME or algorithmic sampling. Yet in general, risk-based sampling methods might violate the prohibition of discrimination as profiles might be critically linked to protected grounds, such as ethnicity or nationality. According to Dutch Equal Treatment Law, selection criteria for profiling need to be objective, proportional and necessary to realize the aim pursued. Assessment of these provisions are, however, a

[6] Discrimination through Risk Profiling, Dutch College of Human Rights 2021 https://open.overheid.nl/documenten/ronl-c409ea31-2c00-4318-9a45-d47ad8a2ca7f/pdf

[7] College voor de Raad van Beroep (CRvB), CRvB Jun-5th 2018, ECLI:NL:CRVB:2018:1541, r.o. 4.2; CRvB Sep-20th 2016, ECLI:NL:CRVB:2016:4160, r.o. 4.4.5; CRvB Apr-14th 2015, ECLI:NL:CRVB:2015:3249, r.o. 4.5

normative and context-dependent exercise. Absent specific jurisprudence on the use of algorithmic profiling by Dutch municipalities, for this case study we cannot yet rely on clear legal rulings[8]. Nonetheless, a normative assessment for the issue at hand is still urgently required. This constitutes the objective of this case study.

The Dutch Work and Assistance Act guarantees a minimum income for all legal inhabitants of The Netherlands who have insufficient means to maintain themselves. In return, the Act formulates certain obligations to social welfare recipients, such as participation in a 'trajectory to work' program. More details on the legal basis that obligates Dutch municipalities to conduct social welfare re-examinations in the context of the Work and Assistance Act can be found in Appendix A – Legal background of social welfare re-examinations. In addition, requirements for municipalities described in the Dutch General Administrative Law Act and the European General Data Protection Regulation when conducting risk-based profiling to sample citizens are also discussed in Appendix A – Legal background of social welfare re-examinations.

Process of re-examination interviews and the concept of fraud

The outcome of a re-examination interview is binary; either social welfare is *duly* or *unduly* granted. Unduly granted payment can be broken down in three categories: 1) reclamation, 2) termination and 3) administrative adjustments. An overview of the outcomes of re-examination interviews conducted in 2017-2019 is displayed in Figure 2. More information on the number of completed re-examination interviews per sampling method can be found in Appendix D – Sampling performance metrics.

Based upon the outcome of a re-examination interview, the municipality can decide to start a follow-up investigation to examine fraud, in which 'deliberate deceit' by the reci-

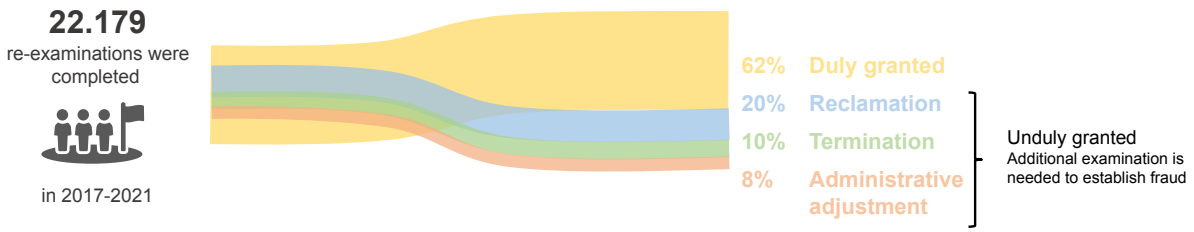

Figure 2 – Outcome of re-examination interviews in the period 2017-2021 conducted by the Municipality of Rotterdam

[8] The Dutch System Risk Indicator (SyRI) court ruling was not based on the workings of the algorithm. The source code of the algorithm has never been made available to the Court. Therefore, Article 8 of the ECHR (respect for private life) was used as a legal 'safety net' to argue for the potential harmful effects of the algorithmic system

pient 'to gain an illegitimate advantage' should be established[9]. The fraud investigation process is considered to be beyond the scope of this case study. It is important to note that due to a certain level of legal discretion, the commitment to trace undue welfare claims varies across Dutch municipalities. Some Dutch municipalities are more 'aggressively' pursuing potentially undue welfare claims, while other municipalities are more lenient.

Issue I: The proxy discrimination and correlation challenge

In the context of social welfare re-examination through algorithmic profiling, the Rotterdam Court of Auditors has flagged a critical link between illiteracy and ethnicity[10]. The Court warned that including literacy rate as a selection criterion for profiling creates a risk of discriminatory bias, since literacy acts as a *proxy variable* for the protected attribute of origin. In their report, however, an assessment of the objectivity, proportionality and necessity of this proxy variable is not given. This is problematic since additional evidential requirements are needed under EU and Dutch non-discrimination law to establish indirect (proxy) discrimination[11]. Following the rationale of the Rotterdam Court of Auditors, *all* used variables are proxies, since every variable used in data modeling is by definition at least partially correlated with a protected group attribute. Simply taking the inevitable statistical occurrence of correlation as a sufficient criterion for indirect (proxy) discrimination, would render the use of all characteristics unjustified and would severely, if not fully, restrict the use of profiling methods as such. This does not seem to be intended by the Rotterdam Court of Auditors, but their own reasoning is insufficient to provide an alternative.

The Court's report lacks a clear qualitative rationale for why, and when, the quantitative issue of correlation with protected attributes becomes intolerable from the perspective of non- discrimination. For some variables, the risk of serving as a proxy for protected attributes might be more tolerated than for others. For instance, because the direct relevance of some variables (say, household composition) for assessing unlawful welfare payments means that their discriminatory risk is weighed differently compared to others (say, literacy rate). In addition, the legal criterion for indirect discrimination – that a variable is 'critically linked' to protected characteristics – is open for qualitative interpretation[11]. The same holds for the criteria for objectivity, proportionality, and necessity to realize the aim pursued. In short, qualitative evaluation is needed to provide an adequate basis for determining the discriminatory character of a profiling method and the selection criteria used. The aim of this case study is to conduct a qualitative assessment of the use of proxy variables: Which variables are, and which variables are not justified to be included in algorithmic sampling and SME sampling in the context of social welfare re-examination?

Dutch municipalities have access to a wide array of data about social welfare recipients.

[9] See Clause 225, 227b and/or 326 of the Dutch Criminal Code

[10] Colored Technology, Rotterdam Court of Auditors 2021 https://rekenkamer.rotterdam.nl/onderzoeken/algoritmes/

[11] S. Wachter, B. Mittelstadt, C. Russell, Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI p.15 (2020). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3547922

For this case study, available selection criteria for the Municipality of Rotterdam are shared. Appendix B – Data collection gives the full list of 60 variables, grouped in 13 categories, which can potentially be used for both SME sampling and algorithmic sampling by the municipality. Out of this list, the final model as implemented by the Municipality selected the top-20 variables for use with the highest predictive power to predict undue welfare claims upon re-examination.

The objective of this case study is not to determine which of the 60 potential variables or which of the 20 used variables are acceptable to be included in risk profiles as selection criteria, yes or no. Rather, it is to establish a reasoning which (types of) variables – considering such aspects as proxy-risk, the extent of their critical link with protected attributes, and their proportionality and necessity for the aim pursued – can be justifiably used.

Figure 3 displays the top-20 variables with most predictive power as indicated by the algorithmic *gradient boosting model* (gbm) profiling method. The essentials of the gbm method are explained in Box 3. Relative importance is a metric that computes the predictive value of input variables (where the variable with most predictive power is assigned 100 by definition). The precise numerical value of this metric is less important than the general magnitude and order of importance among the different variables. A description of the top-20 most predictive variables is given in Appendix C – Variable selection methods.

Among five different candidate models, the gbm algorithm was considered the best performing method by the Research and Business Intelligence team of the Municipality of Rotterdam, which collaborated in this project with consultancy firm Accenture. Subsequently, the gbm variable selection method has been adopted for actual use. That means, complemented by other sampling methods, in 2017, 2018 and 2019 respectively 10%, 17% and 22% of all re- examination interviewees were selected by algorithmic sampling. More statistics on the scale and performance of the discussed profiling and sampling methods are attached in Appendix C – Variable selection methods.

Issue II: Ethically (un)desirable criteria for risk profiling

Aside from the issue of serving as proxy-variables for protected attributes, there are ethical concerns about using certain personal characteristics for profiling as such. In this part of

Box 3

What is a gradient boosting model?

Gradient boosting is a machine learning technique used for classification tasks. It gives a prediction based on an ensemble of hundreds or thousands of decision trees. In terms of accuracy, boosting models usually outperform simpler models, such as single self-explainable decision trees or logistic regression methods, but sacrifice interpretability, since the relative importance of input data variables is based on an aggregation statistic of the total ensemble.

the case study we explore the ethical considerations having to do with potentially unfair, irrelevant, subjective or otherwise problematic personal data used for risk-profiling.

Data points about recipients that are available to many Dutch municipalities include the number and duration of appointments with a job coach, attendance of job seeking events, and how recipients communicate with representatives of the municipality, e.g., by email, telephone or by post. In Rotterdam, additional data on personal characteristics of recipients are collected, such as mental care needs and received psychological help. During a personal interview (not a re-examination interview) trained civil servants score recipients on competences and work skills, e.g., motivation, flexibility, perseverance, self-reliance, pro-activeness, stress resistance, presentation, professional appearance, eagerness to learn, and attitude. The full list of all data points is displayed in Appendix B – Data collection. An example scoring sheet can be accessed online[12].

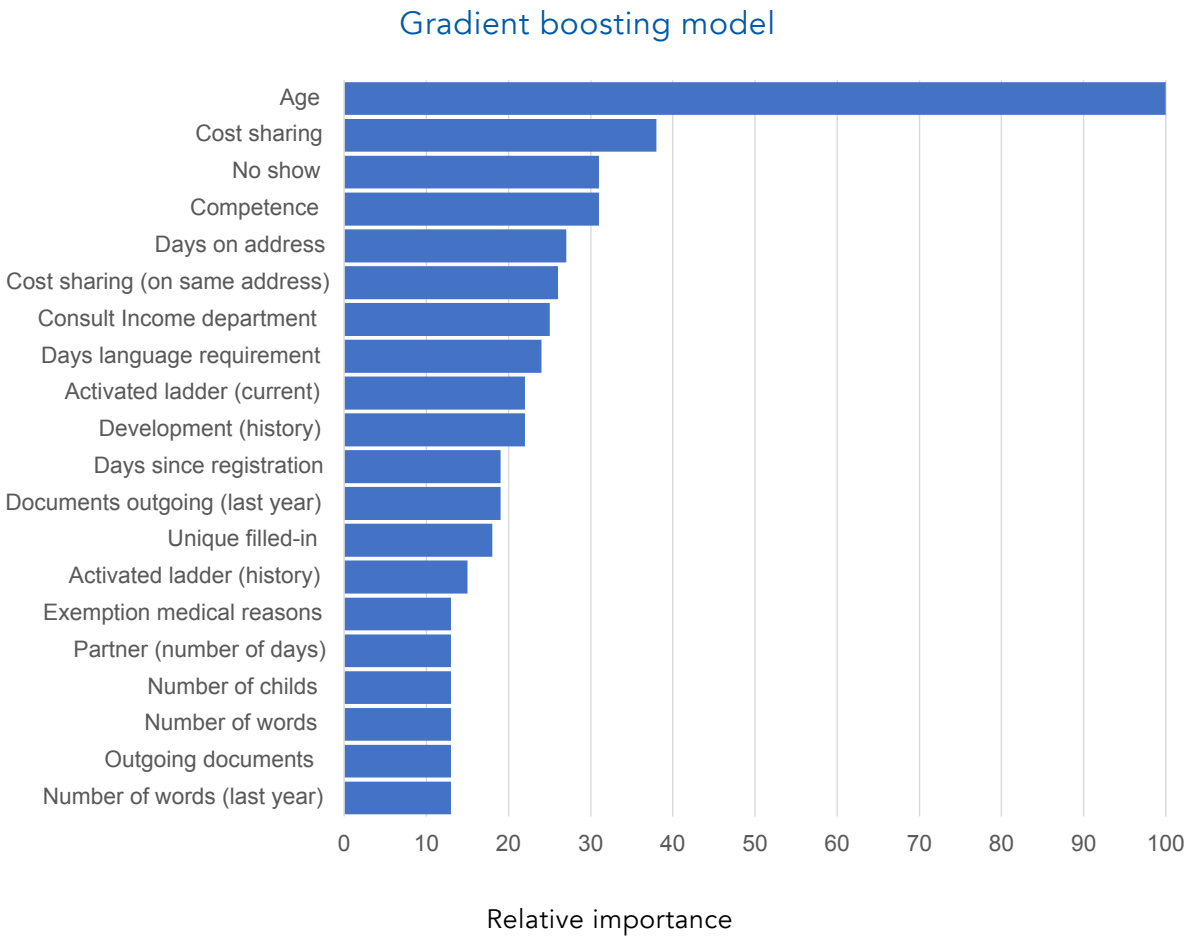

Figure 3 – Top-20 most predictive variables according to the gradient boosting model (gbm) profiling algorithm as deployed by the Municipality of Rotterdam. Predictive power is measured in terms of relative importance (where the variable with most predictive power by definition is assigned score 100). A description of all variables is provided in Appendix B – Data collection.

[12] See https://algorithmaudit.sharepoint.com/:b:/s/CaseRotterdam/Ec6gNE1_PwJOieXYW0rG-oMB2LO2VyYFOj-leuF0xz0Ogg?e=clGa2z

One could question whether some of the characteristics are desirable to serve as selection criteria in risk profiles to select recipients for re-examination. Not because of the risk of proxy- discrimination, but because no objective and reasonable justification for distinctions on the basis of these grounds can be provided. For instance, differentiation based on mental care needs (no classical protected ground in non-discrimination directives), could still be perceived as unfair differentiation, since it could reinforce social inequality. Or it could be questioned whether it is fair to take extraneous traits, such as the recipients' 'professional appearance', or the nature of the conversation between a civil servant and recipient, into account for establishing risk-profiles. Another concern is the subjective nature of some data points, such as the level of 'flexibility or 'autonomy', being subjectively assessed by a civil servant.

On the one hand, it could be of added value for a municipality to be aware that some citizens have a certain potentially vulnerable background. Digitizing and storing this information would enable the municipality to provide adequate help. On the other hand, one should critically reflect on whether those highly personal, sensitive, possibly subjective variables should be digitized in the first place, and secondly made available as selection criteria for risk profiling.

Enerzijds kan het voor een gemeente van toegevoegde waarde zijn om te weten dat sommige burgers een bepaalde achtergrond hebben. Door deze informatie te digitaliseren en op te slaan zou de gemeente adequate hulp kunnen bieden. Anderzijds moet men zich kritisch afvragen of deze zeer persoonlijke, gevoelige, en mogelijk subjectieve variabelen in de eerste plaats moeten worden gedigitaliseerd, en vervolgens beschikbaar moeten worden gesteld als selectiecriteria voor risicoprofilering.

Whether selection criteria for risk profiling are selected through algorithmic variable selection methods or through SME expertise, a relevant question in this context is what characteristics are ethically undesirable to use in profiling methods, for reasons other than discriminatory bias. The second issue of this case study is, hence, to assess what selection criteria can legitimately be included in risk profiles. It is related to the first issue, but now focusing on the ethical concerns about variables as such. This is a contextual normative exercise for which no silver bullet exists.

Issue III: Comparing SME and algorithmic profiling methods

In the third part of this case study, we compare the ethical implications of SME and algorithmic profiling methods to generate risk profiles. Does it matter whether selection criteria and risk profiles are derived by an SME or an algorithm? Including this aspect in this case study is necessary, to avoid an unwarranted prejudice towards algorithmic methods used for variable selection. SME profiling, based on manually chosen selection criteria, is the most direct alternative to algorithmic profiling. Yet it is important to recognize that it still relies on risk profiles that are potentially discriminatory, unfair or stigmatizing. Hence, critical overemphasis on the practice of using machine learning for building risk profiles,

might result in the increased use of manual SME profiling, even if it is unclear whether that actually reduces the ethical risks.

SME checks

For both SME and algorithmic profiling, it is necessary to manually evaluate the variables which are potentially used as selection criteria. Both methods therefore require 'humans in the loop' that check the profiling process and outcome for potential biases. In Rotterdam, this is implemented in the following way. For SME profiling, a risk profile is manually generated by civil servants who typically choose 1-3 selection criteria during an expert meeting. An integral component of such expert meetings is a qualitative evaluation whether selection criteria are relevant, desirable and potentially discriminatory. For example, if civil servants observe a gender disparity in the sample over the year, they may choose to manually include a gender criterion for the following year to counter the disparity. Such human interventions in the profiling and sampling process are called *SME checks*.

SME checks are also part of the algorithmic sampling process. This happens at two distinct points in the process. First, in the stage where the algorithmic model is trained, an SME check is done on the input data and the model. Part of the check in this stage is an assessment of which input variables to exclude on ethical or legal grounds. Regarding the model, SMEs need to decide about the number of selection criteria to be included. In Rotterdam, the decision was made to include the top-20 most predictive variables (as identified by the algorithm) in the final risk profile. Second, once the trained model is applied to generate samples of recipients, a mandatory SME check is carried out on the model predictions, i.e., the sampled population. An SME checks the sample for possible biases or other wrongful characteristics.

Hence, SME checks that need to be carried out are similar for both SME and algorithmic profiling, and involve similar ethical questions. For both SME and algorithmic profiling, civil servants are confronted with the question whether their risk profiles are fair, and which selection variables are ethically justified to use. This entails, as mentioned earlier, that one-sided critique of algorithmic profiling does not solve the fundamental ethical issues, since they also apply to SME profiling.

Differences SME and algorithmic profiling

Nonetheless, there are differences between SME and algorithmic profiling that should be considered. The most obvious difference is the higher accuracy of algorithmic profiling compared to SME profiling. As attested in the case of Rotterdam, in the period 2017-2021 algorithmic sampling systematically outperforms SME sampling in practice to predict unduly granted welfare payments. For this period, the accuracy of algorithmic sampling, conditioned on reclamations, is approximately 29.9% where the accuracy of SME sampling is approximately 16.5% (n=4,388). Higher accuracy not only implies a higher efficiency for detecting unduly granted welfare claims, beneficial to the municipality. It is arguably also beneficial for welfare recipients, as it implies a reduction in 'false positives', i.e., recipients

that are needlessly selected for time-consuming and stressful re-examination of what are in truth lawful welfare claims (see Box 1). Performance metrics for the discussed sampling methods (SME and algorithmic risk profiling, event-driven sampling and random sampling) can be found in Appendix D – Sampling performance metrics. In this appendix, other relevant performance metrics of the algorithmic sampling method are provided as well.

Another relevant difference is the fact that algorithmic methods make use of much more data and personal characteristics. This is clear in the case of Rotterdam, where SME profiling typically uses 1-3 selection variables, whereas the algorithmic model uses 20 variables. Using more variables means using more potentially problematic variables: for 20 variables, it is more likely that it includes discriminatory proxy-variables, or subjective and irrelevant personal characteristics. It also generates more complexity. Algorithmic risk models combine personal characteristics in a statistically sophisticated way. It is often difficult to explain in non-technical, accessible language why exactly an individual recipient has been assigned a high risk by a model. For algorithmic profiling, an explanation for sampling can be provided by means of the relative importance of selection criteria based on the training dataset, as displayed in Figure 3. Yet this is still not very insightful for explaining an individual decision. For SME profiling, this is much easier, as it is simply a matter of whether a recipient matches a pre-defined risk profile. Hence, SME profiling is better explainable than algorithmic profiling. Lastly, the relatively straightforward risk profiles chosen by SMEs make it easier to vary and to diversify profiles over time, in order to break historical biases. For instance, a clear profiling bias with respect to certain neighborhoods in one year, can be opposed by manually choosing to emphasize other neighborhoods in the next year. Although possible, for algorithmic profiling methods, it is more difficult to manually force these variations. These are some notable, ethically relevant differences between algorithmic and SME profiling.

In sum, SME and algorithmic profiling are alternative methods with their own ethical up- and downsides. For a comprehensive evaluation of risk profiling used by Dutch municipalities, weighing SME against algorithmic methods is indispensable. Both are evidently imperfect. Both methods contain irreducible bias and unfairness, and reach limited performance. Yet they do so in a different way, and the question is what in the context of this case is the better alternative – or whether, perhaps, both profiling methods are unjustified.

Appendix A – Legal background of social welfare re-examinations

Eligible population for re-examination

Not all recipients of social welfare are eligible to be selected for re-examination. Recipients satisfying one of the following criteria are exempted:

> Only once in every two-years social welfare recipients can be selected for a re-examination interview;
> Recipients receiving social welfare for less than 6 months are exempted for re-examination;
> Social welfare recipients ≥64 years old are not eligible for re-examination;
> Social welfare recipients without an address or living in care facilities are not eligible for re-examination.

Dutch General Administrative Law

Decision-making processes by Dutch municipalities are subjected to the Dutch General Administrative Law (in Dutch: Algemene wet bestuursrecht/Awb). This legal framework regulates how governmental bodies, including municipalities, can exercise public power:

> Article 2:4 Awb (prohibition of prejudice);
> Article 3:2 - 3:4 clause 1 Awb (duty of care);
> Article 3:47 Awb (motivation and legal certainty).

Unifying these principles with the use of algorithmic variable selection methods is a challenge. On itself, algorithmic selection of variables is not a decision as defined in Awb Article 1:3, as an employee of the municipality decides whether social welfare is (un)duly granted after the re- examination interview. However, variable selection could be seen as part of the duty of care, i.e., careful preparation of this decision. Difficulties in explaining why certain criteria are include in a risk profile can result in a municipality acting 'lawfully' but not 'appropriately'. Additional organizational and legal requirements on how algorithmic profiling methods can align with the duty of care are, however, an open and context-dependent question. This case study aims to contribute to an answer to this question.

Legal rules in Chapter 5 of the Awb form the legal basis for enforcement of social welfare policies. The Participation Law is part on special administrative law (see below).

General Data Protection Regulation

The General Data Protection Regulation (GDPR) regulates the storage and processing of Dutch social welfare recipients. How data processing methods in this case study relate to the requirements as stated in the GDPR can be found in the Privacy Impact Assessment (PIA) as conducted by the Municipality of Rotterdam[13]. Certain GDPR provisions relevant for this case study are stated below:

[13] Freedom of Information (FOI) request VPRO Argos/Lighthouse Reports https://www.vpro.nl/dam/jcr:c87f2d6c-3f9c-4498-9a9c- f3bc5483a437/Downloads%20Model%20Rotterdam.zip

> Article 4 – Profiling definition

Profiling concerns "any form of automated processing of personal data consisting of the of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements".

> Article 5 – Principles relating to processing of personal data

- Purpose limitation: The Municipality of Rotterdam considers adhering to the principle of purpose limitation since the Participation Law obligates municipalities to trace unduly granted social welfare allowances. All input data is collected for sake of compliance to the Participation Law, i.e., personal client info, work history, trajectory to work plan etc.;
- Data minimalization: As stated in the PIA, the Municipality of Rotterdam considers to adhere to the principle of data minimization, since "as little as possible data is collected for this project" and "only the most predictive selection criteria, as indicated by algorithmic profiling, are considered for sampling methods". However, in the same document is stated that: "as much as possible data variables are considered for model training to reveal unexpected patterns".

> Article 6 – Lawfulness of processing

Article 6 paragraph 1 sub e provides the legal basis for processing, based on the necessity for the performance of a task carried out in the public interest. This task arises from the Participation Act (Article 53a, 64) and the Act on the Structuring of the Implementation Organization for Work and Income (in Dutch: Suwi, Article 62);

> Article 9 – Processing of special categories of personal data

In principle, according to clause 1 of this Article, processing health data is prohibited. If exemptions apply to this context is, as this is being written, unknown for us;

> Article 13-15 – Information to be provided where personal data are collected from the data subject/ -have not been obtained from the data subject and Rights of access by the data subject

Article 13(2)f, 14(2)g and 15(1)h state that the data subject has the right to obtain "meaningful information about the logic involved" pertaining to profiling;

> Article 22 – Automated individual decision-making, including profiling

Civil servants take the final decision whether social welfare allowances are (un)duly granted based on an in-person re-examination interview. Profiling methods therefore serve as a sampling method and are therefore not considered as fully automated decision-making and are therefore not regulated by this article (see also the above section on the Dutch General Administrative Law).

Dutch Participation Law

Article 11 of the Dutch Participation Act (Work and Assistance Act) guarantees a minimum income for everyone who is living legally in The Netherlands and who has insufficient means to maintain themselves. When social welfare allowances are received, Article 7 obligates recipients to participate in a 'trajectory to work'. Municipalities assist recipients in this trajectory. Certain groups, for instance single parents with a child up to 5 years old, may request dispensation from this obligation. People are however obliged to attend training courses. Article 53a and 64 provide the legal basis for (re-)examination.

Dutch Law Structure Executive Organizations Work and Income

Article 62 van de Structure Executive Organizations Work and Income (in Dutch: Suwi) regulates the mutual provision of information between implementation organizations and municipalities.

Appendix B – Data collection

Data points on social welfare recipients collected by the Municipality of Rotterdam

Category	Variable	Example data point
Address	Date of arrival, date of departure	
	Dutch ZIP code	First four elements of code
	City district	"Charlois", "City centre", "Zuidwijk"
Appointment	Date	
	Text describing appointment with job coach, client manager or re-examination interviewer etc.	Description of appointment by civil servant, e.g., "Start social welfare", "End date social welfare reached", "Re-examination interview"
	Result of appointment	"Additional info needed", "Participation in social workplace", "Payments terminated"
	Reason for appointment	"Introductory meeting", "Administrational check", "Social welfare terminated"
	Additional system text	"Client is warned verbally", "Research into extension social welfare", "Subscribed to follow-up stage"
Availability	Start date, end date title	
	Availability title	"Available", "Limited availability due to caring <5 years old"
	Availability description	"Heavy physical and psychological issues", "In therapy for chronic psychosis"
Characterization	Start date, end date	
	Characterization type	"Sector transportation operations and logistics", "Refugee", "Profile: lively, Precise and physically"
Contact	Date of documentation	
	Type of contact	"Email", "Conversation", "Letter"
	Subject	"Trajectory to work", "Contribution assessment", "Employment motivation"
Competences	Start date, end date	
	Competence title	"Empathy", "Assertive to take decisions and plan activities", "Resilient to stress"
	Competence description	"ADHD", "strange language errors", "illiterate", "unconvincing attitude", "conversation not possible in Dutch"

Category	Variable	Example data point
Exemption work obligation	Start date, end date	
	Exemption type	"Social grounds", "Temporary exemption work obligation and compensation", "Medical grounds"
Impediments for participation	Start date, end date	
	Impediment category	"Not tech savvy", Inability to generate income", "Psychological issues"
	Impediment description	"Heavy physical and psychological complaints", "In therapy for chronic psychosis"
Instrument	Start date, end date	
	Reason for termination	"Match", "Goal reached", "Outflow to regular work", "Transferred to prematching", "Work/reintegration", "Supporting instruments", "Activation"
Participation in activ-	Start date, end date	
	Reintegration ladder ('trajectory to work' program)	"Activation (civic participation)", "Supporting instruments", "Work/Reintegration"
	Terminated prematurely	"Found paid work", "Malfunctioning", "Moving outside municipality district"
Person	Gender	
	Month of birth Attitude Autono-	
Personal features	Houding	
	Autonomie	
	Assertiviteit	
	Communicatie	
	Discipline	
	Eager to learn	
	Flexibility	
	Hobbies/Sport	
	Job application behavior	
	Language – Dutch reading capacities	
	Language – Ability to speak Dutch	

Category	Variable	Example data point
	Language – Capacity to under-	
	Language – Writing capacities Dutch	
	Language requirements fulfilled (conversation, listening, reading,	
	Motivation	
	Presentation capacities	
	Professional appearance	
	Other comments	
	Record of Arrests and Prosecutions	
	Trajectory to work activation	
	Working during office hours	
	Working outside office hours	
Relationship (relevant for recipient)	Start date, end date	
	Type	"Parent -> child", "landlord -> tenant", "partner -> partner (married)"
		"Child relation", "flatmate", "spouse"
Steunplan	Start date, Proposed date, Sign date,	
	Plan description	"Mediation", "Diagnosis", "Social activation"
	Reason for termination	"Targets achieved", "Support plan rejected by client", "Outflow as self-employed"
	Description of targets	"Outflow to regular work by means of short professional experience related appointments", "Social engagement by participating in events that aim to contribute to society"

Table 1 – Data points collected about social welfare recipients at the Municipality of Rotterdam[3]

Appendix C – Variable selection methods

SME variable selection

At the Municipality of Rotterdam, the composition of SME-driven risk profiles changes annually. Profiles are guided by certain multi-annual profile domains, such as 'household risk profiles', 'age risk profiles' and/or 'city district profiles'. Examples of household risk profiles are "men living alone" and "single female tenants"[14]. In this process, the involved SMEs, supported by the Data Privacy Officer of the Municipality of Rotterdam are responsible for the legitimacy of the established profiles.

Algorithmic variable selection

A description of the top-20 variables holding most predictive power to predict undue social welfare allowances (as displayed in Figure 3) is provided in Table 2. The gradient boosting model (gbm) has been considered as the best performing algorithm among candidate methods for variable selection on training data set: gbm, glmnet, random forest, rpart and xgbtree.

Variable name	Description
Age	Age of social welfare recipient
Cost sharing	Multiple adults share a living, but do not live together. The amount of social welfare support is reduced
No show	Without notice, the recipient did not show up at an appointment with the municipality
Competence	Competence rated by the Employee Insurance Administration (UWV). Information is retrieved from the UWV database
Days on address	Number of days residing at address
Cost sharing (on same address)	Multiple adults live on the same address. The amount of social welfare support is reduced
Consult Income department	Employees of the Income department have been interacting with the social welfare recipient
Days language requirement	Days since started with language requirements
Activation ladder (current)	Method of 'activation' currently used to guide citizens to work. See also variable 'Instrument' in Appendix B – Data collection
Development (history)	Sum of past activities to support social welfare recipients. See also variable 'Support plan' in Appendix B – Data collection

[14] Section 4-3 in Colored Technology, Rotterdam Court of Auditors (2021) https://rekenkamer.rotterdam.nl/onderzoeken/algoritmes/

Variable name	Description
Days since registration	Days since social welfare has been requested
Documents outgoing (last year)	Number of writings last year to social welfare recipients with description 'document outgoing'
Unique filled-in	Sum of reported interactions
Activated ladder (history)	Sum of 'activations' stored in the column 'Reintegration Ladder'. See also variable Reintegration in Appendix B – Data collection
Exemption medical reasons	Number of days that a social welfare recipient has been exempted from 'trajectory of work' in the past due to medical circumstances
Partner (number of days)	Number of days a partner has been registered
Number of childs	Number of (foster) childs registered
Number of words	Sum of the number of words used by civil servants in the field 'text' in the past
Outgoing documents	Number of writings in the past to social welfare recipients with description 'document outgoing'
Number of words (last year)	Sum of the number of words used by civil servants in the field 'text' in the past year

Table 2 – Description of top-20 most predictive variables according to gbmodel

Appendix D – Sampling performance metrics

Quantitative aspects

Quantitative performance metrics shed light into the effectiveness of profiling methods. These metrics are key to make business decisions, but should also inform ethical assessments. For instance: If an algorithmic risk profiling method generates a high gain in accuracy, the ethical risks are weighted differently compared to when the accuracy gain is only minimal. Both quantitative and qualitative aspects should inform the ethical evaluation.

Performance per sampling method in Rotterdam

The Municipality of Rotterdam shared the outcome of re-examination interviews per sampling method for the period 2017-2021. Four different outcomes are considered: no action, termination, reclamation and administrative adjustments. Only the latter three outcomes are displayed in Table 3 - Table 7. The total number of social welfare recipients in the Municipality of Rotterdam in 2017-2021 were: 44.600 (June 2017), 42.110 (June 2018), 9.840 (June 2019), 48.830 (June 2020), 46.090 (June 2021)[15].

2017	Completed re-examinations	Termination		Reclamation		Administrative adjustments	
Algorithmic sampling	255	44	17,3%	65	25,5%	-	0,0%
SME sampling	1.817	317	17,4%	269	14,8%	9	0,5%
Event-driven sampling	166	73	44,0%	32	19,3%	-	0,0%
Random sampling	2	-	0,0%	-	0,0%	-	0,0%
Total	2.240	434	19,4%	366	16,3%	9	0,4%

Table 3 – Accuracy of sampling methods at the Municipality of Rotterdam in 2017

2018	Completed re-examinations	Termination		Reclamation		Administrative adjustments	
Algorithmic sampling	477	86	18,0%	100	21,0%	6	1,3%
SME sampling	2.013	264	13,1%	395	19,6%	22	1,1%
Event-driven sampling	305	116	38,0%	54	17,7%	1	0,3%
Random sampling	2.961	190	6,4%	515	17,4%	35	1,2%
Total	5.756	656	11,4%	1.064	18,5%	64	1,1%

Table 4 – Accuracy of sampling methods at the Municipality of Rotterdam in 2018

[15] Dutch National Office of Statistics https://opendata.cbs.nl/statline/#/CBS/nl/dataset/80794ned/table?dl=34613

2019	Completed re-examinations	Termination		Reclamation		Administrative adjustments	
Algorithmic sampling	1.428	188	13,2%	381	26,7%	18	1,3%
SME sampling	2.946	210	7,1%	612	20,8%	36	1,2%
Event-driven sampling	824	251	30,5%	199	24,2%	14	1,7%
Random sampling	1.169	64	5,5%	278	23,8%	20	1,7%
Total	6.367	713	11,2%	1.470	23,1%	88	1,4%

Table 5 – Accuracy of sampling methods at the Municipality of Rotterdam in 2019

2020	Completed re-examinations	Termination		Reclamation		Administrative adjustments	
Algorithmic sampling	1.207	102	8,5%	431	35,7%	317	26,3%
SME sampling	1.850	67	3,6%	418	22,6%	384	20,8%
Event-driven sampling	265	73	27,5%	101	38,1%	45	17,0%
Random sampling	24	1	4,2%	13	54,2%	8	33,3%
Total	3.346	243	7,3%	963	28,8%	754	22,5%

Table 6 – Accuracy of sampling methods at the Municipality of Rotterdam in 2020

2021	Completed re-examinations	Termination		Reclamation		Administrative adjustments	
Algorithmic sampling	253	12	4,7%	105	41,5%	98	38,7%
SME sampling	3.841	77	2,0%	359	9,3%	388	10,1%
Event-driven sampling	134	48	35,8%	53	39,6%	26	19,4%
Random sampling	242	4	1,7%	8	3,3%	15	6,2%
Total	4.470	141	3,2%	525	11,7%	527	11,8%

Table 7 – Accuracy of sampling methods at the Municipality of Rotterdam in 2021

Accuracy of gradient boosting model on new data

Performance of the gradient boosting model on a control data set, i.e., newly collected data that was not part of the train, validate and test data set on which the model was trained. The data set consists of SME, event-driven and randomly sampled recipients and the outcome of the forthcoming re-examination interview.

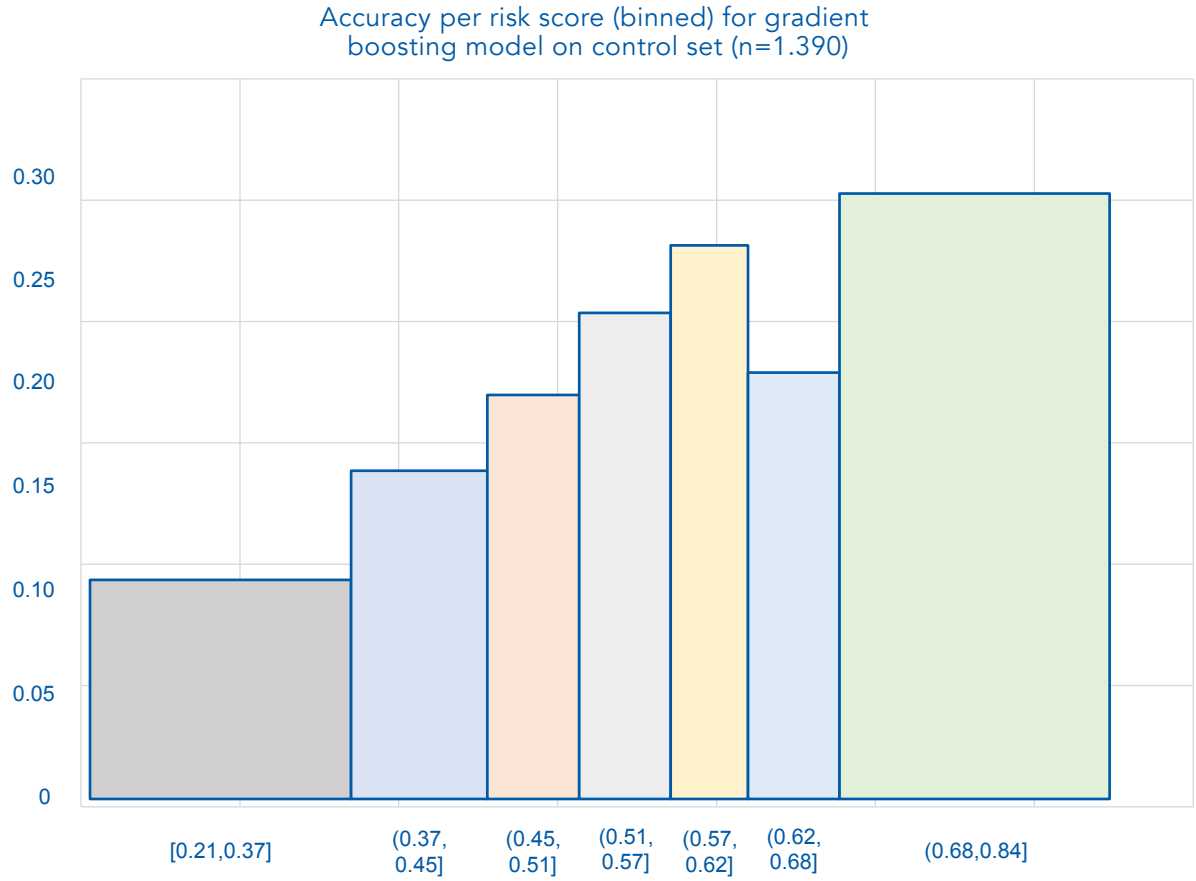

Figure 4 – Accuracy on real-life control data set (n=1.390)

Re-examination statistics from Amsterdam, The Hague and Utrecht

Three of the four largest Dutch municipalities (Amsterdam, The Hague and Utrecht) shared statistics on the number of re-examination interviews and the outcomes (duly or unduly granted). Algorithmic sampling was only used in the Municipality of Rotterdam to sample recipients for re-examination, according to our information. Recipients in other municipalities are selected for re-examination only through SME, event-driven and random sampling. In 2021, approximately 30% of all social welfare recipients in The Netherlands (±365.000) lived in the four largest municipalities (Amsterdam, The Hague, Rotterdam and Utrecht)[16].

Amsterdam

The Municipality of Amsterdam provided statistics on the number of finished re-examinations and the sum of repayments as a consequence of re-examination per year.

[16] Dashboard on social welfare, Dutch National Office of Statistics (CBS) https://dashboards.cbs.nl/v2/dashboardSOZ/

Year	Number of social welfare recipients[16]	Number of re-examinations[17]	Repayments
2021	39.378	1.728	€4.153.540
2020	40.298	2.715	€6.797.687
2019	39.633	2.277	€4.873.928

Table 8 – Statistics on social welfare re-examinations in the Municipality of Amsterdam

The Hague

The Municipality of The Hague provided statistics on the number of re-examinations interviews, the outcomes of the re-examinations (duly or unduly granted) and the sum of repayment per year.

Year	Number of social welfare recipients[16]	Number of finished re-examinations[18]	Outcome	Number	Repayments
2021	24.020	1.698	Duly granted	909	€4.153.540
			Unduly granted	789	
2020	24.683	1.292	Duly granted	685	€6.797.687
			Unduly granted	607	
2019	24.668	1.937	Duly granted	1.289	€4.873.928
			Unduly granted	648	

Table 9 – Statistics on social welfare re-examinations in the Municipality of The Hague

[17] Data provided by the Municipality of Amsterdam upon request

[18] Data provided by the Municipality of The Hague upon request

Utrecht

Every quarter, the Municipality of Utrecht publishes statistics on the number of re-examination interviews and the outcomes (duly or unduly granted).

Year	Number of social welfare recipients[16]	Number of finished re-examinations[19]	Outcome	**Number**
2021	10.460	523	Rechtmatig toegekend	202
			Onrechtmatig toegekend	321
2020	10.695	741	Rechtmatig toegekend	293
			Onrechtmatig toegekend	448
2019	10.518	1.629	Rechtmatig toegekend	611
			Onrechtmatig toegekend	1.018

Table 10 – Statistics on social welfare re-examinations in the Municipality of Utrecht

[19] Data provided by the Municipality of Utrecht in letters to the city council on work and income https://utrecht.bestuurlijkeinformatie.nl/Reports/Item/779c76d3-7fcd-4661-9c99-df75ba58a3b1

Funding of Algorithm Audit

Algorithm Audit is a nonprofit organization supported by independent public funding. We are committed to balanced, careful and independent review of ethical issues that arise in algorithmic use cases. Budget is allocated to draft unsolicited problem statements. We reimburse experts that take part in our audit commissions to carry out the evaluations of ethical issues. We serve society and the international AI auditing community by making all our advice and knowledge public. Working nonprofit suits our activities and goals best.

Structural partners of Algorithm Audit

SIDN Fonds

The SIDN Fund stands for a strong interntet for all. The fund invests in bold projects with added societal value that contribute to a strong internet, strong internet users, or that focus on the internet's significance for public values and society.

European AI&Society Fund

The European AI & Society Fund supports organisations from across Europe that want to shape policies so that Artificial Intelligence better serves people and society. It's a collaborative fund, which means a group of foundations have come together to pool their resources.

Dutch Ministry of Internal Affairs

The Dutch Ministry of Interal Affairs promotes the democratic state, the rule of law and sound public administration. It safeguards core values of democracy. The ministry advances digitalizing public administrations and governmental and public organisations which citizens can trust.

Algorithm Audit

www.algorithmaudit.eu

www.github.com/NGO-Algorithm-Audit

info@algorithmaudit.eu

Stichting Algorithm Audit is registered as a non-profit organisation at
the Dutch Chambre of Commerce under license number 83979212