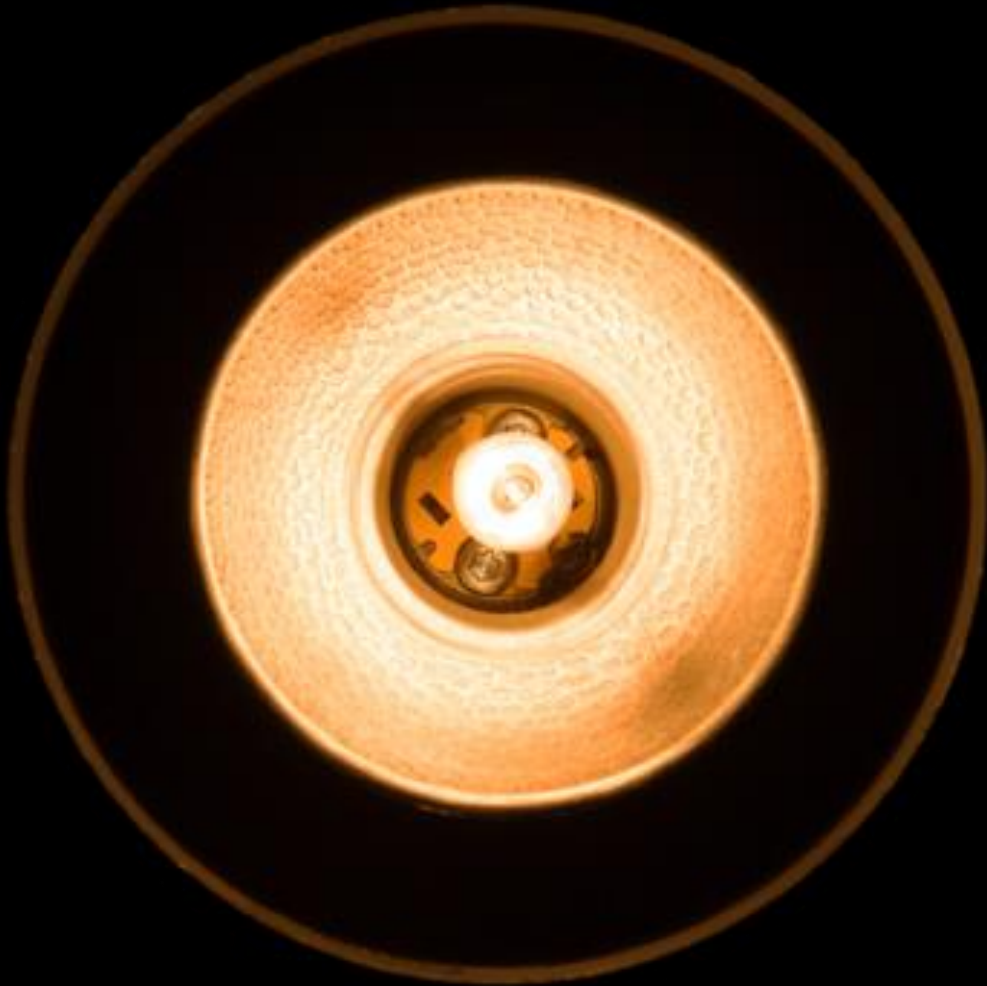


Deloitte.



**Better use of data and advanced
statistics / machine learning in
delivering benefits to the fuel poor**

Deloitte LLP

May 5th 2020

Important notice from Deloitte LLP

This Final Phase 2 Research Report (the "Final Report") has been prepared by Deloitte LLP ("Deloitte") for the Department for Business, Energy and Industrial Strategy (BEIS) in accordance with the contract with them dated 19th December 2019 ("the Contract") and on the basis of the scope and limitations set out below. The Final Report is also provided to the BEIS's partner / sponsored body the Committee on Fuel Poverty (CFP) under the terms of the Beneficiary Access Agreement with them dated 2nd April 2020.

The Final Report has been prepared solely for the purposes of assessing the better use of data and advanced statistics in delivering benefits to the fuel poor, as set out in the Contract. It should not be used for any other purpose or in any other context, and Deloitte accepts no responsibility for its use in either regard, including its use by BEIS or CFP for decision making or reporting to third parties.

The Final Report is provided exclusively for BEIS and CFP's use under the terms of the Contract and Beneficiary Access Agreement respectively. No parties other than BEIS and CFP are entitled to rely on the Final Report for any purpose whatsoever and Deloitte accepts no responsibility or liability or duty of care to any parties other than BEIS and CFP in respect of the Final Report or any of its contents.

The scope of our work has been limited by the time, information and explanations made available to us. The information contained in the Final Report has been obtained from BEIS, CFP and third party sources that are clearly referenced in the appropriate sections of the Final Report. Deloitte has neither sought to corroborate this information nor to review its overall reasonableness. Further, any results from the analysis contained in the Final Report are reliant on the information available at the time of writing the Final Report and should not be relied upon in subsequent periods.

All copyright and other proprietary rights in the Final Report remain the property of Deloitte LLP and any rights not expressly granted in these terms or in the Contract are reserved.

Any decision to invest, conduct business, enter or exit the markets considered in the Final Report should be made solely on independent advice and no information in the Final Report should be relied upon in any way by any third party. This Final Report and its contents do not constitute financial or other professional advice, and specific advice should be sought about your specific circumstances. In particular, the Final Report does not constitute a recommendation or endorsement by Deloitte to invest or participate in, exit, or otherwise use any of the markets or companies referred to in it. To the fullest extent possible, Deloitte, BEIS and CFP disclaim any liability arising out of the use (or non-use) of the Final Report and its contents, including any action or decision taken as a result of such use (or non-use).

Contents

1	Executive summary	4
2	Introduction	7
2.1	Background and context	7
2.2	Research Objectives	9
2.3	Methodology	10
2.4	Limitations	13
2.5	This report	14
3	The application of machine learning to fuel poverty	15
3.1	Typical implementation design	15
3.2	An algorithmic decision-making framework – learnings from UK policing	19
4	Key challenges and mitigations	21
4.1	Overview	21
4.2	Overarching recommendations	21
4.3	Key challenges and mitigations	23
5	Scope for benefits	43
5.1	Overview of potential benefits	43
5.2	Overview of potential trade-offs	48
6	Conclusions and next steps	50
	Appendix 1 – Machine learning terminology	52
A1.1	Defining key concepts	52
A1.2	Why these techniques?	53
A1.3	Example use cases	53
	Appendix 2 – Detailed case study findings	55
A2.1	Overview of case study findings	55
A2.2	United Kingdom – Risk of reoffending	55
A2.3	Portugal – Risk of long-term unemployment	60
A2.4	Switzerland – Refugee resettlement	65
A2.5	United States – Risk of becoming homeless	68
	Appendix 3 – The ALGO-CARE framework	71
A3.1	The ALGO-CARE framework – learnings from UK policing	71
	Appendix 4 – Estimation of benefits	73
A4.1	Benefit estimation – worked example	73
	Glossary	78
	Bibliography	80

1 Executive summary

Current policies aimed at assisting the fuel poor are constrained in their ability to target fuel poor households. This report analyses the proposed use of machine learning techniques to improve the identification of these households, including the scope for benefits, challenges to implementation and potential mitigations.

A person is defined as being in fuel poverty under the Warm Homes and Energy Conservation Act 2000 if they are a member of a household living on a lower income in a home which cannot be kept warm at a reasonable cost.¹ Under the current Low Income High Cost measure, 2.5 million households in England are estimated to be fuel poor, representing 10.9% of the total.²

In 2014, the Government put in place a new statutory fuel poverty target for England, with the objective to ensure that as many fuel poor households as reasonably practicable achieve a minimum energy efficiency rating of Energy Performance Certificate (EPC) Band C by 2030.³

A number of support schemes can provide either bill reductions or energy efficiency installations to fuel poor households. Nevertheless, a key challenge to delivering these schemes has been the identification of households for support, many of whom may be unaware that they meet the eligibility criteria. Energy suppliers, for example, incur significant search costs in seeking to deliver energy efficiency measures under the Energy Company Obligation (ECO). The receipt of benefits is often used as a proxy to determine eligibility, however only c.49% of fuel poor households were in receipt of benefits as of 2017.⁴

The Department for Business, Energy and Industrial Strategy (BEIS) and its partner organisation the Committee on Fuel Poverty (CFP) are therefore researching the possible implementation of machine learning techniques that could improve the identification of fuel poor households; in particular for schemes such as ECO and the Warm Home Discount (WHD) Broader Group which currently require customers to apply, or be identified, for support. The application of these techniques to the ECO scheme would likely be of highest priority, given the higher search costs incurred compared to other schemes. Improved targeting could lead not only to the more effective delivery of support measures but also reduce the cost of delivering the Fuel Poverty Strategy.

The implementation of machine learning techniques should be considered as a set of processes and methodologies; documenting the considerations made towards these at an early stage can allow these to be refined and improved over time. The objective of this research is therefore to assess the potential use of machine learning techniques to improve the identification of fuel poor households; this includes identifying the key challenges to implementation and associated mitigations, analysing

¹ Warm Homes and Energy Conservation Act 2000. Available at:

<http://www.legislation.gov.uk/ukpga/2000/31/section/1>

² Fuel Poverty Statistics 2019 (2017 data) Available at:

<https://www.gov.uk/government/collections/fuel-poverty-statistics>

³ BEIS (2015): Cutting the cost of keeping warm. Available at:

<https://www.gov.uk/government/publications/cutting-the-cost-of-keeping-warm>

⁴ Fuel Poverty Statistics 2019 (2017 data) Available at:

<https://www.gov.uk/government/collections/fuel-poverty-statistics>. Includes households that report receipt of means-tested benefits/tax credits, Attendance Allowance, Disability Living Allowance or Personal Independence Payment.

the potential scope for benefits over and above existing policy measures, and delivering evidence-based recommendations to inform future phases of work. To inform the research, four case studies were identified whereby machine learning or statistical techniques have been implemented in a social policy context.

The key findings from this research are as follows:

- **There are a number of relevant use cases whereby machine learning, automation or other statistical methods have been applied successfully in a public and social policy context**, often in cases where the final decision has potentially material consequences for those affected. Case studies in the UK, USA, Portugal and Switzerland⁵ identified significant improvements through the use of machine learning or statistical methods in reducing the risk of crime through reoffending, reducing the likelihood of long term unemployment, reducing long term homelessness and placing refugees in locations to optimise employment prospects.⁶
- **There is a large potential scope for benefits from the implementation of machine learning techniques**, over and above existing policy measures. Improved identification can not only improve the provision of targeted support but also lead to wider positive externalities on the environment, health and wellbeing. An assumption-based illustrative example in this report suggests that improvements in the identification of fuel poor households for ECO support would not only enable energy and bill savings for this group, but also enable a net benefit for society as benefits are diverted to those most in need, even if existing scheme resources were held constant. Assuming 25% of fuel poor households in England currently ineligible for ECO support could be identified through machine learning, the example implies that total social benefits of c.£10m per year could be achieved. The example also highlights considerable scope for reductions in search cost from current levels.
- **Although there exists a wide range of potential implementation challenges, evidence does not suggest that any one given challenge represents a barrier that cannot be mitigated to at least some extent** within a prospective implementation, assuming the model passes acceptability thresholds for predictive power.
- **Amongst the challenges identified, ethical challenges regarding potential data and algorithmic biases represent some of the largest risks to implementation.** It should be ensured that a robust framework is in place to monitor the ongoing performance of the algorithm and mitigate potential discriminatory outcomes, together with biases in the underlying data. The latter is of particular importance; underlying biases can be exacerbated as the model is updated to incorporate prior decisions, which could further increase the risk of discriminatory outcomes over time. Indeed under the Public Sector Equality Duty,⁷ public bodies are required to have due regard to the objectives of the Equality Act (2010),⁸ in particular to eliminate discrimination and advance equality of opportunity.
- **Caution should also be applied in seeking to deliver fully automated solutions through machine learning.** Machine learning techniques may not be able to capture the full range of features that determine the target variable, and are inherently unable to avoid false negative predictions that could lead to the exclusion of particular subgroups. At a minimum, a suitable challenge mechanism should be developed whereby non-recipients can re-open their individual cases for review, for example if pursuing automation of the Warm Home Discount Broader Group.

⁵ The case studies chosen for this report were derived from a long list and selected based on policy relevance.

⁶ A Guide to using Artificial Intelligence in the Public Sector (2020). Available at: <https://www.gov.uk/government/publications/a-guide-to-using-artificial-intelligence-in-the-public-sector>

⁷ The Public Sector Equality Duty (2011). Available at: <https://www.gov.uk/government/publications/public-sector-equality-duty>

⁸ Equality Act (2010). Available at: <http://www.legislation.gov.uk/ukpga/2010/15/contents>

This report provides a number of overarching recommendations to inform the future implementation of machine learning techniques. These include:

- i) Developing a clear governance framework** to ensure a standard process for risk mitigation, ownership and accountability within the machine learning implementation. This would cover all stages of the implementation from data collection and processing through to the day-to-day usage and ongoing monitoring of the algorithm.
- ii) Considering the acceptability thresholds** for adopting machine learning tools to identify fuel poor households, in particular regarding false negatives and predictive power. This is not only a statistical consideration; seeking to automate a model with inaccuracies could not only exclude eligible households but also incorrectly identify households as eligible for support, which could, in turn, inflate the cost of delivering the schemes.
- iii) Determining how best to operationalise model outputs** in arriving at a policy decision. In particular, a set of business rules would need to be developed that define how model outputs translate into tangible actions with regards to the targeting of fuel poor households, together with the level of human involvement. For example, if households are identified as fuel poor with high likelihood, business rules would determine what the next steps would be regarding verification or the offer of support.
- iv) Conducting an ethical and legal impact assessment** to assess whether the proposed use of the tool is in line with the relevant legal and ethical requirements. This could include the completion of a Data Protection Impact Assessment (DPIA), which seeks to identify and mitigate any particular risks regarding the usage of personal data, but should also include a wider assessment of the potential implications for equality and human rights. Legal advice should be sought to assess alignment with, for example, GDPR, the Public Sector Equality Duty and Human Rights legislation within the proposed implementation design.

Further suggested research could update or widen the selection of case studies identified in this study, develop a formal framework for cost-benefit analysis or assess the factors that influence the uptake of support schemes in greater depth. Importantly, machine learning can only act as an enabler in identifying fuel poor households; if the uptake rate of support schemes remains low this may considerably reduce the scope for benefits from implementation.

2 Introduction

The 2015 Fuel Poverty Strategy for England, "Cutting the cost of keeping warm", aims to reduce bills and increase wellbeing in the coldest, lowest income homes. This report provides evidence-based recommendations for the proposed use of machine learning to better identify these households and improve the provision of targeted support.

2.1 Background and context

Under the Warm Homes and Energy Conservation Act 2000, a person is defined as being in fuel poverty if they are a member of a household living on a lower income in a home which cannot be kept warm at a reasonable cost.⁹ According to the 2019 Fuel Poverty Statistics (2017 data), 10.9% of all households in England were living in fuel poverty under the Low Income High Cost (LIHC) measure.^{10,11}

In 2014, the Government put in place a new statutory fuel poverty target for England, with the objective to ensure that as many fuel poor households as reasonably practicable achieve a minimum energy efficiency rating of Band E by 2020, Band D by 2025 and Band C by 2030. The 2015 Fuel Poverty Strategy for England subsequently set out a vision for meeting the fuel poverty target, with the ambition to reduce bills and increase wellbeing in the coldest, lowest income homes.¹²

The Department for Business, Energy and Industrial Strategy has implemented a range of schemes designed to support those in need and reduce the number of fuel poor households.¹³ Some provide financial assistance, whereas others offer the physical installation of energy efficiency measures to drive cost reductions and improve household energy efficiency. These schemes include:

- *Warm Home Discount (WHD)* - provides vulnerable consumers in Great Britain with a £140 rebate on their energy bill each year. Eligibility for the scheme is either through receipt of the Guarantee Credit element of Pension Credit (the "Core Group"), or if a customer is on a low income and receives certain means-tested benefits (the "Broader Group"). Discounts are automated for the Core Group, whereas those eligible for the Broader Group are required to apply directly to their supplier.

⁹ Warm Homes and Energy Conservation Act 2000. Available at: <http://www.legislation.gov.uk/ukpga/2000/31/section/1>

¹⁰ The LIHC measure classifies a household as being fuel poor if its fuel costs are above the average and its disposable income (after housing and fuel costs) is below the poverty line. BEIS is currently consulting on changing the definition of fuel poverty to the "Low Income Low Energy Efficiency" (LILEE) measure. This would define a household as fuel poor if disposable income (after meeting required housing and energy costs) is below the poverty line and if the property has an EPC rating of Band D or lower. The introduction of the LILEE measure would increase the number of households considered fuel poor from 2.5 to 3.6 million in England (2017 data). The Consultation on the Fuel Poverty Strategy for England (2019), which proposed the use of LILEE, is available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/819606/fuel-poverty-strategy-england-consultation.pdf

¹¹ Fuel Poverty Statistics 2019 (2017 data) Available at: <https://www.gov.uk/government/collections/fuel-poverty-statistics>

¹² BEIS (2015): Cutting the cost of keeping warm. Available at: <https://www.gov.uk/government/publications/cutting-the-cost-of-keeping-warm>

¹³ BEIS website: <https://www.gov.uk/government/organisations/department-for-business-energy-and-industrial-strategy>

- *Energy Company Obligation (ECO)* - enables households to reduce the cost of heating their property through the installation of energy efficiency measures (e.g. wall insulation). Eligibility is determined through receipt of certain qualifying benefits. Each obligated supplier has an overall target for support provided, based on its share of the domestic energy market. Customers are required to apply directly to, or can be identified by, their participating supplier. Under ECO3, Local Authorities are also able to widen the eligibility criteria in order to tailor support measures to their area, often based upon income or health considerations (the "LA Flex" mechanism).
- *Winter Fuel Payment (WFP)* - a universal pensioner benefit based on age and residence. Pensioners are entitled between £100 and £300 a year depending on their circumstances, and receive WFP automatically. If a customer is eligible but does not get paid automatically, they are able to make a claim and can request a mandatory reconsideration of the outcome.
- *Cold Weather Payment (CWP)* - a payment of £25 for each 7 day period of very cold weather (zero degrees Celsius or below) between 1st of the November and the 31st of March. Eligibility is determined through receipt of certain benefits, with payments made automatically.
- *Minimum Energy Efficiency Standard for Landlords* - sets a minimum efficiency standard of EPC Band E for landlords, subject to a cost cap of £3,500 including VAT per property.

The Committee on Fuel Poverty, a partner organisation sponsored by BEIS, was launched in 2016 to advise on the effectiveness of these policies and encourage greater coordination across the organisations working to address fuel poverty.^{14,15}

In 2017, 92.2% of fuel poor households lived in a property with an energy efficiency rating of Band E or above, with corresponding figures of 65.9% and 10.0% for Bands D and C respectively. This compares favourably with 81.1%, 32.7% and 1.5% observed in 2010.¹⁶ The average fuel poverty gap, which measures the reduction in fuel bills that the average fuel poor household needs in order to not be classed as fuel poor, also decreased to £321 in 2017 from £333 one year earlier.¹⁷

Despite progress against the target, a key challenge in delivering these measures to the fuel poor has been in the identification of households for support, many of whom may be unaware they meet the eligibility criteria. This is particularly the case for ECO and the WHD Broader Group, given the requirement for those eligible to apply directly (or be identified by suppliers / Local Authorities within ECO). The receipt of certain benefits is often used as a proxy to determine scheme eligibility, however only c.49% of fuel poor households were in receipt of benefits as of 2017.^{18,19}

As a result, current targeting efficiency for programmes aimed at assisting the fuel poor ranges from less than 10% for Winter Fuel Payment to 30% for the Energy Company Obligation (i.e. only c.30%

¹⁴ CFP website: <https://www.gov.uk/government/organisations/committee-on-fuel-poverty>

¹⁵ CFP Framework Document (2019). Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/806568/cfp-framework-document-2019.pdf

¹⁶ BEIS (2019) - Annual Fuel Poverty Statistics 2019, 2017 data. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/829006/Annual_Fuel_Poverty_Statistics_Report_2019_2017_data.pdf

¹⁷ BEIS (2019) - Fuel Poverty Factsheet. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/808300/Fuel_poverty_factsheet_2019_2017_data.pdf

¹⁸ BEIS (2019) - Annual Fuel Poverty Statistics 2019, 2017 data. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/829006/Annual_Fuel_Poverty_Statistics_Report_2019_2017_data.pdf. Includes households that report receipt of means-tested benefits/tax credits, Attendance Allowance, Disability Living Allowance or Personal Independence Payment.

¹⁹ Policies such as ECO have however been successful at providing support to eligible households across Great Britain. Statistics published in April 2020 show that since the start of ECO3 in October 2018, around 2.7m measures have been installed in 2.1m households up to the end of February 2020. The statistics are available at: <https://www.gov.uk/government/statistics/household-energy-efficiency-statistics-headline-release-april-2020>

of ECO3 energy efficiency measures will be installed in fuel poor homes).²⁰ Improvements in the identification of fuel poor households could lead not only to the better provision of targeted support but also lower the cost of delivering the strategy through reductions in search costs. This could subsequently inform the design of future schemes to more effectively target fuel poor households.

2.2 Research Objectives

BEIS and its partner organisation CFP are seeking to study the better use of advanced statistics and machine learning (ML) in delivering benefits to the fuel poor; these techniques can improve the identification of fuel poor households to enable the more effective delivery of targeted support. In particular, the study assesses how the potential barriers and challenges associated with the implementation of these techniques can be overcome.

The objective of this research is to develop an evidence-based framework for implementing these methods, drawing on the learnings from other countries. To support the research, four case studies are presented whereby machine learning or other forms of automation have been implemented in a social policy setting.

The objectives of this report are summarised below:

- Understand BEIS's current capabilities regarding the potential use of machine learning tools to identify fuel poor households;
- Develop four case studies where government departments have implemented machine learning methods in a public policy setting, conducted through an analysis of publically available material and stakeholder interviews;
- Identify the key challenges to using machine learning and assigning benefits automatically, in particular regarding the legal and ethical considerations;
- Understand how the challenges associated with the implementation of machine learning may be mitigated, drawing on the learnings from the case studies and wider research;
- Evaluate the scope for benefits from implementing machine learning and artificial intelligence methods compared to the systems that were previously in place; and
- Transfer the key learnings from this research into recommendations for BEIS to effectively implement ethical and legally compliant policies.

More widely, the findings from this report may be used by CFP to make recommendations as to how advanced statistics or machine learning could be used to improve the delivery of the Fuel Poverty Strategy. The ultimate objective is to inform the potential implementation of machine learning, which could in turn:

- Accelerate BEIS's work on advanced statistics and the application of machine learning techniques;
- Inform the design of future fuel poor household energy efficiency schemes;
- Reduce the cost of delivering targeted support measures through improved identification; and
- Further the debate regarding the use of machine learning methods in policy implementation and automated benefit assignment.

²⁰ CFP (2018) - Third Annual Report: Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/754361/Committee_on_Fuel_Poverty_Annual_Report_2018.pdf

2.3 Methodology

The research for this report has been conducted in two distinct phases, comprising both independent research and stakeholder interviews:

2.3.1 Phase 1 – Research context and case study selection

2.3.1.1 Research methodology

Interviews with BEIS and CFP stakeholders were conducted to develop a better understanding of the Fuel Poverty Strategy, together with the overall research objectives. In particular, interviews with policy advisors at the Warm Home Discount, Energy Company Obligation, BEIS's analytical team and CFP members gathered insights on the current fuel poverty policies, the operation of the schemes as well as BEIS's ongoing research into data-driven tools.

Together with these interviews, Phase 1 research identified a long list of potential case studies where government departments across the world have implemented machine learning or other artificial intelligence techniques in a public policy setting. The case studies were designed to complement the research by providing real-world examples of the implementation process, the challenges faced and mitigations adopted.

Each of the identified case studies were assessed against four selection criteria to shortlist four case studies for this research report:

- **Policy relevance:** This criterion assessed whether a policy has similar objectives to the fuel poverty case, in particular if the policy assigns benefits or support measures to vulnerable individuals.
- **Data framework:** A case study scored highly against the data framework criterion if the country has a similar privacy and legal framework to England, whereby policies have to comply with data protection regulations similar to GDPR²¹ and the Digital Economy Act.²²
- **Methodology:** Case studies scored highly if they used relevant machine learning, artificial intelligence techniques or automated benefit assignment. A case study also scored highly if the government department used several datasets to train the model and could provide insights on the practicalities and legal requirements of data matching.
- **Information availability:** The information availability criterion provided an indication of the amount of detailed publically available information for each case study. This also assessed the availability of government stakeholders for interview in Phase 2 of the research.

While all of these criteria are of importance, the main focus has been on policy relevance when selecting four case studies for this research report.

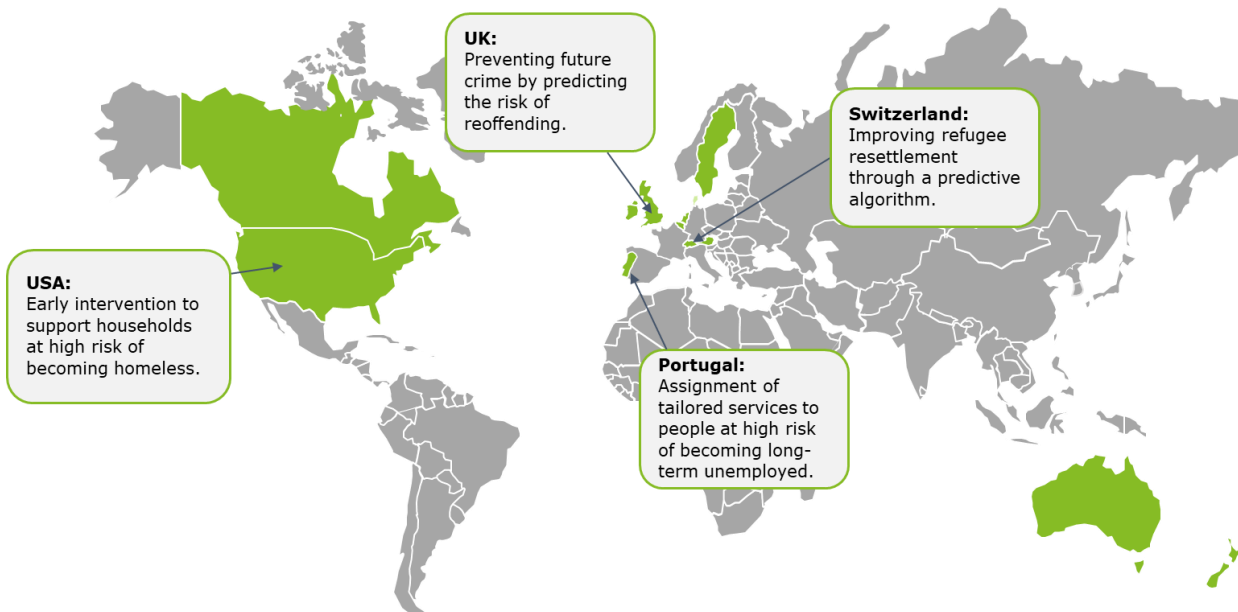
2.3.1.2 Case studies selected

The four case studies selected for Phase 2 are outlined in Figure 1 and summarised below. Also shown with green shading are the countries where further case studies were identified but not selected.

²¹ General Data Protection Regulation (GDPR). Available at: <https://gdpr-info.eu/>

²² Digital Economy Act (2017). Available at: <http://www.legislation.gov.uk/ukpga/2017/30/contents/enacted>

Figure 1: Overview of all identified case studies, including a short description of those selected for this report



Source: Deloitte analysis of publically available information

The four case studies selected for Phase 2 of the research are summarised in further detail below. Further detailed information is set out in Appendix 2.

Table 1: Case studies selected for this report

Country	Use case	Summary
United Kingdom	Risk of reoffending	<p>The Durham Constabulary currently operates the “Checkpoint programme”, a rehabilitation scheme that seeks to reduce the chance of future criminal offences by providing tailored support to offenders.</p> <p>In order to identify people who are eligible for the programme, the Constabulary developed a decision-support tool, namely the Harm Assessment Risk Tool (HART). HART was implemented in 2017 and helps to inform a decision as to whether suspects are at low, moderate or high risk of reoffending within two years, and hence their eligibility for the scheme.</p> <p>The model includes information on 104,000 custody events over a five year period (2008-2012). It uses 34 predictors to arrive at a forecast, most of which focus on the offender’s history of criminal behaviour.</p>
Portugal	Risk of long-term unemployment	<p>The Portuguese Institute of Unemployment and Professional Training (IEFP) has implemented a predictive tool that forecasts a citizen’s risk of becoming long-term unemployed, to help address the high long-term unemployment rates observed in the country.</p> <p>Since 2018, IEFP has explored the use of the tool to assign preventive services such as skill development to people that are categorised to be at high risk of long-term unemployment.</p> <p>The ensemble approach²³ (gradient boosted trees) uses IEFP data on candidate background, records regarding their interaction (e.g. training courses) with IEFP as well as sociodemographic information. The dataset currently includes 3.5 million people recorded between 2007 and 2017.</p>

²³ An "ensemble" approach uses multiple learning algorithms to obtain better predictive performance.

Country	Use case	Summary
Switzerland	Refugee Resettlement	<p>The Swiss State Secretariat for Migration (SEM), in partnership with the Immigration Policy Lab (IPL) at ETH Zurich and Stanford University, developed an algorithm that seeks to allocate refugees to regions where they are most likely to find employment, and hence improve integration. Currently, refugees are allocated to cantons through random assignment.</p> <p>In order to find the optimal resettlement location for each refugee and their families, the algorithm predicts the probability of employment at each of the 26 Swiss cantons (regions) based upon a number of socioeconomic characteristics. A case worker then is responsible for making the final decision on refugee assignment, in part using the outputs of the model.</p> <p>The predictive tool has been trialled in Switzerland since 2018 as a decision-support tool providing placement officers with a suggested optimal location. The tool draws on data of 22,159 refugees (arriving in the period 1999-2013) from the ZEMIS database, collected by SEM.</p>
USA	Risk of homelessness	<p>The New York City Department of Homeless Services (DHS), in partnership with community non-profit organisations, introduced the HomeBase programme in 2004 to improve homelessness prevention. Responsibility for the scheme is now held by the Human Resources Administration (HRA).</p> <p>HRA currently implements a statistical scoring system to assess the risk of becoming homeless. Data obtained from interviews with families seeking support is combined with data on eviction, shelter history and benefits to determine the level of support provided.</p> <p>Households receive HomeBase services based on their risk score; full support includes emergency rental assistance, access to job training and landlord mediation services. The programme provides support to c.28,000 families each year.</p> <p>HRA is currently planning to implement machine learning techniques to improve the outreach of the scheme.</p>

2.3.2 Phase 2 – Additional research and interviews with case study stakeholders

In Phase 2, stakeholders across the four selected case studies were identified and contacted for interviews to complement the research. The objective of these interviews was to develop a further understanding of the schemes and how the key challenges to implementation were mitigated.

Potential interviewees were identified based on their level of involvement in the policy, including stakeholders across the following areas:

- **Research and data scientists:** Interviews with researchers and data scientists were held to gain insights on data management and the methodology used to develop the machine learning algorithm or statistical method.
- **Policy advisors:** Interviews with policy advisors were conducted to obtain detailed information on the machine learning implementation process, the key challenges and possible mitigations.
- **Ethical and legal advisors:** Interviews were held to develop an understanding of the legal and ethical challenges associated with the implementation of machine learning methods and the steps that were required to ensure that policies were legally and ethically compliant.

In order to impose a degree of consistency across the case studies, a set of questions was designed and agreed with BEIS and CFP prior to conducting the interviews.

These interview questions covered the following areas:

- **Background:** explored the overall motivations and objectives of the policy in question, for example the basis for implementing machine learning over existing methods.
- **Methodology and data:** identified the data sources that were used to develop the algorithm, together with any requirements for matching data across multiple sources. Moreover, these questions sought to understand the statistical techniques employed and the accuracy of these techniques over and above existing methods.
- **Implementation and communication:** obtained information on the implementation steps and the requirements to introduce a policy that uses outcomes from machine learning. These questions also explored how the case study engaged with both citizens and “case workers” who use the algorithm on a daily basis.
- **Key challenges:** examined the key challenges to implementation and how these were mitigated. These included practical challenges, legal issues such as GDPR compliance together with ethical considerations regarding fairness, biases and potential discrimination.
- **Next steps:** developed an understanding of how the policy in question is being implemented, together with any further research being conducted in this policy area.

In total seven interviews with stakeholders were conducted. These interviews, together with publically available information, subsequently informed the challenges, mitigations and recommendations set out in Section 4.

2.4 Limitations

This research report is subject to the following limitations, which should be considered in conjunction with the findings presented:

- **Time and resources:** This research report has been conducted over a three month period. Further research over a greater period of time and including a wider range of evidence may have delivered additional and potentially different insights to those set out in this report.
- **Case study selection:** A long list of case studies was identified through desktop research; four were selected for this research report based on four selection criteria agreed with BEIS and CFP, with a main focus on policy relevance. Different selection criteria may have led to a different choice of case studies. Future analysis may also uncover a more recent set of country examples than those presented in this report.
- **Stakeholder interviews:** Interviews with case study stakeholders were conducted over a three week period. The interview process was highly dependent on the availability and willingness of these stakeholders. In total, 24 stakeholders were contacted with seven interviews conducted, covering all of the selected case studies.
- **Legal and ethical framework:** The legal and ethical frameworks applicable to machine learning implementation are continually evolving. This report is only able to raise potentially relevant considerations based on publically available information at the time of writing.

This report does not constitute legal advice; whilst the report highlights a number of current frameworks based on publically available information, it cannot and does not provide legal advice on the application of these frameworks. As stated throughout this report, it is recommended that a full legal review is undertaken ahead of the prospective implementation of machine learning models. For example, legal advice should be sought to assess alignment with GDPR,²⁴ the Public Sector Equality Duty²⁵ and Human Rights legislation, amongst others.

²⁴ General Data Protection Regulation (GDPR). Available at: <https://gdpr-info.eu/>

²⁵ The Public Sector Equality Duty (2011). Available at: <https://www.gov.uk/government/publications/public-sector-equality-duty>

Guidance such as the UK Government guide to using AI in the public sector should also be considered.²⁶

- **Challenges to implementation:** This document focuses on the implementation challenges of machine learning methods that are most relevant to the fuel poverty case, and does not seek to present an exhaustive list of challenges faced within the general field of artificial intelligence or machine learning methods.

2.5 This report

The remainder of this research report is structured as follows:

- Section 3 provides further context regarding the use of machine learning methods to better identify fuel poor households.
- Section 4 sets out the key challenges and mitigations associated with the implementation of machine learning in a public policy context. The section also provides overarching recommendations that should be considered within the prospective implementation.
- Section 5 evaluates the scope for benefits from implementing machine learning techniques, meanwhile recognising the potential trade-offs.
- Section 6 draws together the conclusions of this research report and sets out possible next steps.

²⁶ A Guide to using Artificial Intelligence in the Public Sector (2020). Available at: <https://www.gov.uk/government/publications/a-guide-to-using-artificial-intelligence-in-the-public-sector>

3 The application of machine learning to fuel poverty

This section sets out a generalised framework for the application of machine learning methods. It presents a number of considerations that should be made both prior to and throughout implementation, drawing upon an algorithmic decision-making framework established in the UK policing context.

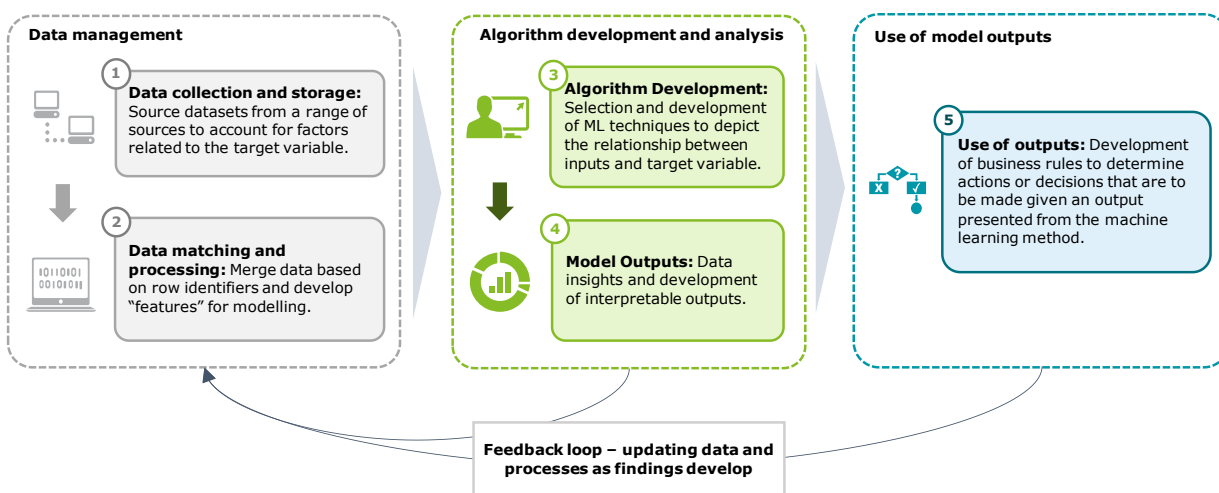
3.1 Typical implementation design

3.1.1 High level implementation stages

At a high level, the stages involved in applying a machine learning tool²⁷ can be represented by the diagram in Figure 2.²⁸ Any implementation of machine learning methods should be an iterative process that seeks to account for the best available data using the appropriate statistical techniques, meanwhile remaining conscious of the legal and ethical requirements.

Each implementation problem will have its own nuances dependent on both the context and complexity of the issue at hand. The chosen design mechanism should be informed at least in part by the policy objectives and desired outputs. As important as the development of the machine learning tool is how outputs are to be used; if this is not clearly defined, the scope for benefits from the tool will be inherently limited.²⁹

Figure 2: Generalised machine learning implementation steps



Source: Deloitte analysis

²⁷ Appendix 1 presents further detail on machine learning methods and their real-world application.

²⁸ It is recognised that the graphic in Figure 2 is a highly simplified representation; the purpose of this section is to introduce the key considerations and stages that may comprise a future implementation framework.

²⁹ The Government Digital Service (GDS) and the Office for Artificial Intelligence (OAI) have further published guidance on how to build and use AI in the public sector. Available at: <https://www.gov.uk/government/collections/a-guide-to-using-artificial-intelligence-in-the-public-sector>

Data management

The accuracy and appropriateness of a machine learning model will be dependent on the coverage and quality of data that is used to train, test and validate it. In general terms, the data management function might comprise:

- 1. Data collection and storage:** the identification and collection of inputs relevant to the target variable or for the creation of model features. This step may include the collection of datasets from multiple sources; selection should consider data limitations such as biases. Whilst data gathering is an important step, there may be diminishing returns to predictive power if significant features are already accounted for within existing datasets. Data privacy requirements such as the personal data minimisation principle under GDPR also require consideration at this stage.³⁰
- 2. Data matching and processing:** Data matching based on household or individual-level indicators would be required when sourcing from multiple databases, to form the modelled dataset. A matched dataset should be evaluated for possible biases or inaccuracies. This stage also involves the development of features for modelling (i.e. factors related to the target variable).

Algorithm development and analysis

There are a number of possible machine learning methods which can be selected to model the target variable as a function of modelled features; further detail is provided in Appendix 1. In developing the algorithm, consideration should be made as to the policy objectives and purpose of implementing these techniques:

- 3. Algorithm Development:** Machine learning methods are selected based on an assessment of statistical criteria and considering the required qualities of the model. Typically, studies that implement machine learning techniques often utilise a combination of algorithmic approaches to improve predictive power. A variety of statistical “accuracy” criteria exists; their applicability or weight will likely be dependent on the context. For example, in some settings it might be pertinent to minimise false negative predictions, over and above general accuracy.³¹

In developing an algorithmic assessment, at a minimum data should be divided into training and test sets; the former to develop and refine the model, the latter to evaluate the performance of the model across an “unmodelled” population.³²

Machine learning algorithms are typically compared to traditional methods such as logistic regression to determine the improvement in precision or accuracy; it is not always the case that machine learning algorithms will outperform these techniques. There is often a common trade-off between model accuracy and interpretability; more complex machine learning methods, particularly unsupervised techniques, may produce outputs that cannot easily be explained.

- 4. Model outputs:** The use of outputs from the model should be clearly considered, for example how a predicted probability is translated into a risk scorecard or recommendation. Outputs can be presented dependent on the policy objectives, for example as risk classifications or a range of probabilities. Model results should also be assessed against ethical considerations such as biases, explainability and discrimination.

³⁰ GDPR Article 5. Available at: <https://gdpr-info.eu/art-5-gdpr/>

³¹ “False negatives” occur whereby a household is classified as not fuel poor but is, in reality, fuel poor. On the other hand, false positives refer to the incorrect classification of a non-fuel poor household as fuel poor.

³² Typically, studies may also separate a “validation” set used to tune model parameters.

Use of model outputs and feedback loops

Having tested the algorithm and determined the appropriateness of model outputs, a decision is required in a policy setting as to how these translate into an action from those using the model, and how the model is to be updated periodically:

- 5. Use of outputs:** Model owners should assess how model outputs are to be used in practice and the level of human oversight required, in line with policy objectives. Business rules should be developed to determine the actions or support measures that would then be selected based on model outcomes. It should also be considered how model outputs can be used consistently by end-users, together with the potential for “automation bias” or similar.³³

To realise the benefits from machine learning, the process outlined above should be iterative and dynamic – that is, as new data is made available and predictions are made, the model is continually updated and tuned based upon these new pieces of information.

These “feedback loops” are a key component of any machine learning implementation and the continued development of the tool. This process should be subject to ongoing monitoring and evaluation; feedback loops could serve to exacerbate underlying data or algorithmic biases.

3.1.2 Application to the fuel poverty case – initial considerations

BEIS has already started to research the use of data-driven tools to enable the improved identification of fuel poor households. A 2017 study employed a supervised machine learning approach (random forest) to predict fuel poor households in England using data from the National Energy Efficiency Dataset (NEED), Experian, the Department for Work and Pensions (DWP), the Ordnance Survey (OS), Valuation Office Agency (VOA) and the English Housing Survey.³⁴

BEIS is continuing to develop these machine learning methods and assess their performance at identifying fuel poor households; this exercise will inform whether these methods are to be considered for implementation in subsequent phases. As such, the formal implementation design and governance framework are still under consideration; this report makes several recommendations in this regard.

Within existing policy measures, machine learning methods would likely be of most use for the ECO and WHD schemes, given other schemes have stricter, measurable criteria and are already mostly automated:

- **Energy Company Obligation:** machine learning could assist with the identification of fuel poor households that are eligible for energy efficiency installations and improve the delivery of targeted support; energy suppliers currently incur significant search costs in this regard. Indeed the final stage impact assessment for ECO3 estimates supplier search costs at a present value of £257m over the period 2018-2022, with sensitivities as high as £1,000 per “lead”.³⁵ Improved identification could facilitate reductions in these costs to deliver the scheme both for suppliers and potentially Local Authorities under the LA Flex mechanism.
- **Warm Home Discount:** Although the provision of discounts to the WHD Core Group is already automated, the Broader Group are required to apply directly to their energy supplier

³³ A phenomenon whereby case workers develop a tendency to rely on the automated suggestion resulting from the algorithm.

³⁴ BEIS (2017). Available at:

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/633228/need-framework-annex-a-fuel-poverty-targeting.pdf

³⁵ ECO3 Final Stage Impact Assessment. Available at:

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/749638/ECO3_Final_Stage_IA_Final.pdf

to receive support. Approximately 50% of fuel poor households (1.2m) were eligible for this group as of 2017. Improved identification through machine learning could therefore improve the allocation of support to eligible fuel poor households that do not currently apply.

It is envisaged that if machine learning methods were to be implemented, any solution would fit within existing policy delivery frameworks. Under current policies, each scheme³⁶ is designated an individual “policy lead” responsible for the overall implementation of the scheme. Given the different criteria and characteristics of each scheme, an option under consideration is to allow for flexibility in how policy leads can utilise outputs from machine learning to improve the delivery of targeted support. How the implementation framework will ultimately be designed is to be assessed in future phases of work.

The below subsections provide a brief overview of the considerations made to date by BEIS against the key implementation steps outlined previously.

Data management

Within previous research, BEIS collected data from NEED, DWP, Ordnance Survey, Experian, the Valuation Office Agency and the English Housing Survey. BEIS’s current work involves not only updating these sources but searching for additional datasets which could improve predictive power, for example, data from HMRC or Local Authorities: these potential additional sources are outlined in blue below. In order to incorporate different datasets into the model, data matching based on household identifiers would be required.

Responsibility for the collection, processing and management of data inputs would be likely conducted by analysts in BEIS. In designing the implementation framework, consideration should be given to how these sources will be cross-examined for potential biases or other data weaknesses.

Table 2: Existing (green) and potential (blue) data sources

Source:	Data:	Source:	Data:
National Energy Efficiency Data	Gas and electricity consumption data and information on energy efficiency installations	HMRC	Income tax data to match to household inhabitants
Department of Work and Pensions	Benefit claimant counts	Local Authority / Council	Potential information on inhabitants and e.g. council tax payments
Ordnance Survey	Building footprint, height and type	Department for Health and Social Care	The Department has a healthcare needs map which may provide a proxy for fuel poverty and energy need.
Experian	Household income, occupiers, tenure at property	Supplier data	Energy consumption data
English Housing Survey	Fuel poverty status		
Valuation Office Agency	Property information such as age, type and floor area		

Source: BEIS, Deloitte analysis

³⁶ These fuel poverty schemes are introduced in section 2.1. Further scheme details are set out on the UK Government and Ofgem websites.

Algorithm development and analysis

BEIS's data science team would be responsible for the development of machine learning algorithms to identify fuel poor households. The 2017 study found that in seeking to reduce the number of false negatives, the model produced a relatively larger number of false positives; of those predicted to be fuel poor in the study only c.25% of these actually were. BEIS is currently seeking to adapt the supervised learning techniques used in this study to improve predictive power.

In determining whether a model is suitable for implementation, BEIS would next need to consider their level of tolerance levels for false negatives or positives within any machine learning solution, and whether this would be justifiable if implemented.

Use of model outputs and feedback loops

A set of actions would need to be defined once a household is categorised as fuel poor, in particular how outputs are used across the different fuel poverty schemes. The frequency and process by which models are updated, together with any ongoing review processes, would also need to be considered.

There are a number of potential options by which the better identification of fuel poor households could be utilised to improve the delivery of support schemes. One model could include working with energy suppliers to target those households identified as fuel poor with the highest probability by the machine learning tool (in seeking to provide ECO support). Whilst this would still involve some administrative costs, identifying these households in the first instance could reduce search costs from current levels. For both ECO and WHD, the insights from machine learning could also be used in future to tailor the design of the schemes and associated eligibility criteria.

As set out above, the implementation design is to be developed further should BEIS pursue the adoption of a machine learning solution. This could also include the creation of a centralised role to coordinate across internal stakeholders. A formal governance framework that incorporates these considerations would need to be developed in future phases of work.

3.2 An algorithmic decision-making framework – learnings from UK policing

The considerations that need to be made at each step of a machine learning implementation are potentially numerous and wide-ranging. The challenge is then to develop a transparent, adaptable decision-making framework that provides practical steps to address them. This section presents such a framework for the deployment of algorithmic assessment tools, drawing upon the "ALGO-CARE" mechanism developed by Oswald et al. (2018) in a UK policing context.³⁷

Policing is a particularly pertinent example for the fuel poverty case; decisions made with the assistance of an algorithm can directly impact those affected in a material way. Moreover, given that the use of statistical tools and techniques in policing has been ongoing for a number of years,³⁸ a considerable amount of work has already been conducted to develop appropriate governance and monitoring frameworks.

Box 1 provides an overview of the background and objectives of the framework and illustrates further the ALGO-CARE mnemonic.³⁹ It is recommended that this framework is considered in conjunction with existing research (e.g. from the British Academy and Royal Society), together with work being undertaken by, for example, the Centre for Data Ethics and Innovation. In particular, the mnemonic provides a useful and transferable resource with which to develop practical steps for machine learning implementations across the public sector.

³⁷ The full name of the framework is "Algorithms in Policing – Take ALGO-CARE™", referred to simply as "ALGO-CARE" within this report.

³⁸ Babuta and Oswald (2020) note research indicating the use of predictive policing methods as far back as 2004.

³⁹ This report reproduces the ALGO-CARE mnemonic and key questions as set out in Oswald et al (2018) pp245-248 in Appendix 3; ALGO-CARE is subject to trademark. Full credit for this mnemonic and the key questions lies with the authors of the original framework. The original should always be referred to when seeking to assess the ALGO-CARE framework in its own right.

A direct quotation of the ALGO-CARE mnemonic and key questions developed by Oswald et al. (2018) can be found in Appendix 3.

Box 1: The ALGO-CARE framework

Case study - ALGO-CARE

The ALGO-CARE framework developed by Oswald et al. (2018) reflects the experience of the Durham Constabulary in developing the HART tool, which seeks to predict the risk of reoffending within two years. The HART tool represents one of the four case studies considered in this research.

The framework aims to translate key legal and ethical challenges into practical considerations and guidance that can be addressed by public sector bodies. The ALGO-CARE framework has been adopted by the National Police Chiefs' Council's (NPCC) Business Change Council, which, together with supporting documentation, recommends its use to Chief Constables. This example brings together the key considerations into a machine learning decision-making framework that is easily transferable to other use cases.

ALGO-CARE mnemonic

- Advisory** – The algorithm should be used in an advisory capacity, with a human (officer) retaining decision-making discretion.
- Lawful** – The (policing) purpose should justify the use of the algorithm, and potential interferences with the privacy of an individual need to be proportionate to the purpose.
- Granularity** – The algorithm should make suggestions at a sufficient level of detail, given the purpose of the algorithm and the nature of the data processed.
- Ownership** – A "senior person" should own the algorithm and the data analysed, be responsible in maintaining and updating the model and make sure that operations are kept secure.
- Challengeable** – Post-implementation oversights and audit mechanisms (e.g. to identify biases) need to be in place.
- Accuracy** – The percentage of false positives / negatives should have an acceptable threshold and model specifications should match the (policing) aim and policy objective. Periodic validation of the stated accuracy is required.
- Responsible** – The algorithm should be considered as fair, transparent and accountable and be placed under review (alongside other IT developments in policing).
- Explainable** – Appropriate information about the decision-making rule(s) and the impact that each factor has on the final score or outcome should be available.

Source: Deloitte analysis of Oswald et al. (2018) pp245-248

4 Key challenges and mitigations

This section discusses the potential challenges involved with the implementation of machine learning techniques to identify fuel poor households, together with associated mitigations. Several overarching recommendations are also provided to inform the implementation design.

4.1 Overview

This section sets out recommendations to inform the implementation of machine learning techniques, together with how the most pertinent practical, ethical and legal challenges may be mitigated.

These recommendations, challenges and mitigations have been developed through an analysis of publically available information, complemented by the research and interviews conducted across the four selected case studies. These case studies are introduced in Section 2.1 and discussed in detail within Appendix 2.

4.2 Overarching recommendations

This study has identified a number of overarching recommendations for consideration within the implementation of machine learning techniques:

1. Governance framework:

The establishment of a clear governance framework is essential to standardise internal processes, ensure accountability and mitigate potential risks within the machine learning implementation.

Should trials of machine learning methods demonstrate sufficient predictive power to be continued further, the BEIS should establish a formal governance framework for the implementation. Coordination is likely to be required across internal teams including policy leads, data owners and data scientists.

Drawing upon both wider research and the learnings from the case studies, it is recommended that this should include, but not necessarily be limited to:

- A standardised process for the documentation of considerations made prior to implementation of the algorithm, drawing on models such as ALGO-CARE (referenced in Section 3.2) and existing work being undertaken by organisations such as the Centre for Data Ethics and Innovation (CDEI).⁴⁰
- A data governance process to manage the availability, quality and security of data collected for modelling purposes, in line with data protection regulations such as GDPR.

⁴⁰ Centre for Data Ethics and Innovation website: <https://www.gov.uk/government/organisations/centre-for-data-ethics-and-innovation>

- A clear division of roles and responsibilities to ensure accountability. This could include the creation of “senior responsible owners” for each stage of implementation, together with an overall process owner that coordinates across these owners.
- The potential development of a challenge mechanism that allows non-recipients to reopen the assessment of their eligibility, should decisions be made directly based on model outcomes.
- A formal ethical and legal review process whereby the proposed and ongoing use of the model and its underlying data can be appropriately scrutinised and validated by independent stakeholders. Development of this framework could include engagement with advisory bodies such as the CDEI to facilitate consistency across government departments.
- Periodic technical reviews whereby the model can be evaluated against its statistical performance and adjusted accordingly.

2. Thresholds for model acceptability:

It is recommended that acceptability thresholds for the adoption of machine learning tools should be developed, particularly with regards to false negative outcomes, together with standard statistical performance metrics. Performance against these thresholds would determine whether the models currently under development should be taken forward into a formal implementation.

In developing these thresholds, the relevant statistical criteria should be informed by the specific policy objectives. For example, minimising false negatives within the optimisation would seek to reduce the number of eligible fuel poor households excluded by the algorithm. This choice may however result in an increase in false positives, which could have wider implications for the delivery of support measures. In particular, a large degree of false positives could reduce the scope for cost efficiencies, particularly if the delivery model involves human input to validate the households identified by the model.

Lastly, consideration should be made as to whether the model has been able to capture a sufficient breadth of available characteristics that determine in practice whether a household is fuel poor. If known relevant factors cannot be considered, this may not only reduce the potential scope for benefits but also the feasibility of implementing the machine learning solution.

3. Operationalisation:

Within the implementation design, consideration should be made as to how the algorithm is operationalised as a practical tool to improve the support provided. Across the case studies, how the results of a tool were presented and used to inform policy outcomes was often observed to be a key determinant in the success of the implementation. As Babuta and Oswald (2020) state, “the decision-making process informed by the algorithm requires as much attention as the tool itself”.

There are a number of considerations that need to be made in this instance:

- **Business rules:** the development of a formal, standardised process that sets out how model outcomes are translated into tangible actions (i.e. the next steps that occur should a household be classified as fuel poor by the algorithm). For example, if households are identified as fuel poor with high likelihood, what the next steps would be regarding verification or notification of their eligibility for ECO support through suppliers.
- **Level of human intervention:** within the business rules above, the desired and feasible level of automation needs to be determined within the implementation. Evidence from the case studies suggests the use of a human-centric decision-making process where model outcomes are used to support and inform decision-making. In practice, this would imply some form of human verification within the decision to offer ECO support. Moreover, under the LA Flex mechanism, Local Authorities could potentially use model outputs to inform their

selection of households and tailor support measures to their area. There is greater potential for automation within the WHD Broader Group (given this represents financial assistance only, with stricter criteria), however, at a minimum, a challenge mechanism would be required whereby non-recipients could re-open their claims.

- **Interface design:** A tool interface that facilitates engagement with those using the algorithms and minimises the risk of automation bias should be developed. Across the HART tool and Portuguese IEFP / Nova research, considerable work has been undertaken to research how case workers react to particular tool designs, including “nudges” and other psychological impacts.
- **Implementation phases:** Should the algorithm meet the necessary performance criteria, it would need to be considered how the tool will be introduced in practice. For example, a phased “shadow” implementation could be considered, whereby the operation of the algorithm is trialled on the most recent data in parallel to existing policies. Across the case studies, considerable time was taken to test and validate model performance on updated data prior to implementation. In the case of refugee assignment in Switzerland, the randomised control trial is to last at least 2 to 3 years to review effectiveness.

4. Impact assessment:

Before the full implementation of a machine learning solution, an impact assessment should be conducted to assess whether the use of the tool is in line with legal and ethical requirements. It is recommended that legal input is sought at the earliest stage within the implementation to ensure alignment with the relevant legal frameworks.

Drawing on the learnings from Babuta and Oswald (2020), this impact assessment should include, but not necessarily be limited to:

- Data Protection Impact Assessment – Identifies and seeks to mitigate any data protection risks regarding the usage of personal data in the machine learning model.
- Equality Impact Assessment – Assesses the potential impact of the project on equality, taking into account protected characteristics (for example race and age); this would include a consideration of the requirements under the Equality Act and Public Sector Equality Duty.
- Human Rights Assessment – Examines any potential conflict with human rights principles, in particular, ECHR Article 8 “right to respect for private and family life”.
- Empirical evaluation of accuracy and operational assessment of real-world effectiveness.
- Assessment of the expected level of errors, and their potential consequences.
- Independent ethical assessment – Identifies any potential ethical consequences regarding the security, rights and wellbeing of individuals affected by model outcomes.

4.3 Key challenges and mitigations

The implementation stages outlined in Section 3 are likely to encounter a number of cross-cutting challenges, which can be broadly categorised into:

- **Practical:** for example regarding the effectiveness of the techniques employed, together with the governance and oversight of the machine learning implementation process.
- **Ethical:** there are wide-ranging ethical considerations to be made across any machine learning implementation. For example, whether an algorithm introduces or perpetuates

existing biases against certain subgroups or the potential implications of false negatives and the resulting impacts on those affected.

Legal: in this context, legal challenges could relate to requirements under GDPR and the use of personal data, for example.

For each category in turn, this section sets out the potential challenges together with recommended mitigations, drawing on the learnings from wider research and the case study examples.

4.3.1 Key challenges identified

While all of the challenges outlined in this section are relevant to the prospective implementation of machine learning techniques, evidence collected within this research suggests that the challenges outlined in Table 3 are likely to be of the highest importance.

Ethical challenges regarding the presence of false negatives, together with potential data and algorithmic biases, are likely to represent some of the largest risks to implementation. In practice, the extent to which each is encountered will ultimately depend upon the implementation design.

Further detail on these challenges and potential mitigations is provided throughout the following sections.

Table 3: Key challenges identified

Category	Challenge	Challenge detail	Summary mitigations
Practical	Predictive Power	<i>Although unlikely, it may not be possible to develop a machine learning model that can accurately identify fuel poor households and deliver benefits over and above existing policies.</i>	<ul style="list-style-type: none"> Predictive power can be improved through the inclusion and regular updating of additional data sources. Performance thresholds should be developed to inform a decision on whether to adopt machine learning methods, given policy objectives and the proposed use of the tool.
Ethical	Presence of false negatives	<i>It is not practically feasible to produce a machine learning model that avoids the presence of false negatives, namely where a fuel poor household is classified as not fuel poor by the machine learning algorithm.</i>	<ul style="list-style-type: none"> Consideration should be made as to what would represent an acceptable threshold for false negatives and positives. These thresholds should also consider wider policy implications, for example improving identification within existing budget constraints. If measures are to be allocated directly based upon model outputs (e.g. for the WHD Broader Group), a suitable challenge mechanism should be developed such that non-recipients can reopen their cases.
Ethical	Automation based on incomplete information	<i>Machine learning models are often limited by data availability and cannot possibly assess all factors (especially those qualitative in nature) that may affect the outcome.</i>	<ul style="list-style-type: none"> It is advised, where applicable, that machine learning be supported by human judgement in seeking to allocate support. Within ECO for example, some form of human verification would likely be required before outputs can be translated into an offer of support. Alternatively, Local Authorities could use insights from machine learning to inform, but not explicitly determine, their selection of households within the LA Flex mechanism.
Ethical	Data biases	<i>There is an inherent possibility in any machine learning implementation that underlying datasets could include biases or not be fully representative of the population under consideration.</i>	<ul style="list-style-type: none"> It would be pertinent to adopt a formal ethical review process whereby underlying datasets are subject to a review for potential bias. Should available data underweight known populations, subsampling or re-weighting may be required.
Ethical	Algorithmic bias and discrimination	<i>Algorithmic bias describes systematic and repeated errors that create unfair outcomes, including privileging certain users over others; these can be exacerbated through feedback loops once prior decisions are incorporated.</i>	<ul style="list-style-type: none"> Ongoing model validation assessments should be conducted to identify potential or ongoing biases occurring within the modelling, once prior decisions are incorporated.

4.3.2 Practical challenges

A number of practical challenges may be encountered within the implementation of machine learning methods, for example regarding the optimisation of the algorithm and the operationalisation of model outputs. The following table sets out these key challenges, together with potential mitigations, that BEIS may wish to consider ahead of the implementation. This section has been informed both through independent research and the key learnings from the four selected case studies, including stakeholder interviews.

Table 4: Practical challenges and mitigations

Challenge	Challenge detail	Recommended mitigation(s)
Optimisation	<p><i>The effective implementation of predictive methods requires a policy decision on what the algorithm should optimise and in what granularity outputs are produced.</i></p> <p>In the fuel poverty case, a pertinent question relates to the potential trade-off between coverage and the cost of the scheme. An algorithm whereby the optimality criterion is set to minimise false negatives (i.e. increasing coverage and inclusion) may result in a larger proportion of false positives, which would potentially increase the cost of the scheme.</p> <p>A number of different questions could also potentially be answered by the machine learning method – usage in this instance would seek to classify households as fuel poor or not fuel poor, however, the tool could also be used to predict the risk that a household may become fuel poor at some point in the future in order to develop pre-emptive remedies.</p> <p>For example, in the Swiss case of seeking to improve refugee allocation, optimisation currently considers the “likelihood of finding a job” over a certain period of time as a proxy for societal integration, however, a number of other criteria such as potential earnings or welfare indicators could also be selected and would need to be tested over time.</p>	<ul style="list-style-type: none"> • The appropriate optimisation criteria of the machine learning solution should be considered, for example minimising false negatives to limit the number of eligible households excluded by the algorithm. • Scheme eligibility criteria could also be considered within the modelling exercise, given that some fuel poor households may not be currently eligible for support within existing policies. In the Swiss refugee allocation case study, constraints such as the maximum number of refugees per canton were programmed into the model to account for these directly.
Predictive power	<p><i>The predictive power of each machine learning model is highly dependent on the underlying data and features included – it may not always be possible to develop a model that suitably captures the underlying inputs.</i></p> <p>This may especially be the case if the factors influencing whether a household becomes fuel poor are in future different from those observed in historical relationships (the “Lucas critique”).</p>	<ul style="list-style-type: none"> • The predictive power of the model may be increased by including, and regularly updating, additional data sources such as income data from HMRC or supplier data, while at the same time taking into account data minimisation principles (e.g. GDPR). Potential data sources that BEIS may wish to consider are listed in section 3.1.2. • As set out within the overarching recommendations, BEIS should consider their tolerance for model inaccuracies

Challenge	Challenge detail	Recommended mitigation(s)
	<p>Limited predictive power (in terms of accuracy and/or precision) may, in turn, limit the use case for the model for public policy purposes.</p> <p>In the 2017 BEIS study,⁴¹ of those who were predicted to be fuel poor, only a quarter of these actually were, indicating a high degree of false positives. Whilst the model was more accurate for the not fuel poor population (over 98% of “not fuel poor” predictions were correct), an overall precision of 23.5% suggests that additional data or other solutions may be required to improve upon this study for use in a public policy setting.</p> <p>The predictive power of the model could have wider policy implications. Not only could the presence of false negatives lead to the exclusion of eligible households, but a model with a high false positive rate could lead to an increased cost of delivering the schemes, particularly if decisions are automated.</p>	<p>or wider performance given their policy objectives – at what point should a model be deemed “implementable”?</p>
<p>Data sharing</p>	<p><i>Limited data availability and data sharing restrictions have the potential to limit the accuracy of the model, and it may not be possible to capture all relevant factors as a result.</i></p> <p>Data sharing may be limited in some circumstances not only due to formal restrictions but also due to cultural differences, data quality or risk aversion. Data sharing and gathering can be both time-consuming and resource-intensive to complete.</p> <p>This concern was also identified across the case studies researched. The HART tool was only able to utilise data held within Durham Constabulary systems, as opposed to other police areas or national IT systems such as the Police National Database. As set out in Oswald et al (2018), not being able to capture all relevant factors represents one reason why the resulting machine learning method can serve only to inform human decision-making, not replace it.</p> <p>The Swiss, Portuguese and US case studies were also similarly limited to relying upon data already held within the relevant government department, even if there was a desire to expand the number of sources considered.</p>	<ul style="list-style-type: none"> • BEIS should consider engagement with other government departments to understand the scope for data sharing, in line with data protection legislation and principles such as data minimisation. • For example, BEIS could consider engagement with bodies such as the Centre for Data Ethics and Innovation (CDEI), an advisory body seeking to connect policymakers, industry, civil society, and the public to develop the right governance regime for data-driven technologies (part of the Department for Digital, Culture, Media and Sport). • Data sharing may require time and is sometimes limited due to data complexity, architecture or data protection regulations. If the desired data cannot be ascertained, BEIS should determine if it is possible to create a model including data already held by the Department within an acceptable accuracy threshold.

⁴¹ BEIS (2017). Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/633228/need-framework-annex-a-fuel-poverty-targeting.pdf

Challenge	Challenge detail	Recommended mitigation(s)
Business rules	<p><i>The effective implementation of a predictive tool requires the introduction of business rules that indicate how model outcomes are translated into a policy action.</i></p> <p>In this instance, BEIS would need to consider how the outputs of the model would be used to inform decisions across any current or future fuel poverty schemes (particularly ECO and WHD) which have varying eligibility criteria and design structures.</p> <p>Within ECO, for example, business rules could determine how the households identified as fuel poor by the model (with the highest likelihood) would then inform the process of suppliers offering support, including any need for verification of whether model classifications hold true in practice.</p> <p>The development of business rules was identified as a key implementation step within the case study research. In the case of predicting the risk of homelessness in New York City, a clear scoring system has been developed that determines eligibility for support. When a risk score reaches a specific threshold, an individual or family is eligible for full HomeBase support measures, with more limited support and advice provided to those not reaching the threshold.</p>	<ul style="list-style-type: none"> • Consideration should be made as to how the outputs of the algorithm would translate into tangible actions across the different fuel poverty policies, particularly ECO and the WHD Broader Group. • Responsible roles and ownership at each implementation stage should be established within an overall governance framework; this could include a centralised role to coordinate model operations across stakeholders.
Technical validation	<p><i>Within the implementation of machine learning methods, a key challenge has been how best to validate the effectiveness of these techniques in practice.</i></p> <p>In many cases, a Randomised Control Trial (RCT) can be used to directly compare the effects on the treated and untreated populations, as currently being conducted for the next 2-3 years by the Swiss Secretariat for Migration (SEM) within the refugee allocation context. In the fuel poverty instance however, it may be challenging to pursue such methods given exclusion and fairness considerations.</p> <p>A challenge is therefore how best to assess and validate both the methodology selected and the effectiveness of these methods. In the 2017 study,⁴² BEIS used English housing survey data on a relatively smaller sample of households to validate findings. A similar exercise is likely to be required in practice together with a process for ongoing monitoring of algorithm performance.</p>	<ul style="list-style-type: none"> • A formal process for the technical review of the methodology selected should be developed, together with the ongoing effectiveness of the machine learning tool. • Within the model development process, model testing could include further validation against the English Housing Survey (as performed in the BEIS 2017 study), using standard techniques. Further validation could include the development of challenger models to act as a comparative measure for model performance, based on alternate techniques (e.g. logistic regression) or specifications. • A “shadow” implementation should be considered, which would involve the “implementation” of the tool in parallel with existing policy measures to validate the model over a defined period.

⁴² BEIS (2017). Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/633228/need-framework-annex-a-fuel-poverty-targeting.pdf

Challenge	Challenge detail	Recommended mitigation(s)
Uptake	<p><i>Machine learning models may be able to improve the identification of those in need. However, there is no guarantee that this results in an increase in uptake of support for policies which depend on agreement with the individual.</i></p> <p>For example, in the case of ECO, individuals may determine that they do not want home insulation measures to be installed, even if this is shown to be for their benefit. Current ECO3 'findability' rates by measure (i.e. the proportion of technical potential that the supply chain can identify and install in any single year) range from c.11% for solid wall insulation up to 100% for central heating measures.⁴³</p> <p>In the case of the New York City HomeBase programme, the scheme has trialled several different awareness campaigns to raise uptake and awareness, such as mobile vans and targeted letters. This illustrated a need to combine predictive analytics with effective practical strategies to raise awareness and encourage uptake.</p>	<ul style="list-style-type: none"> • The development of supplementary models to assess the characteristics of households that accept practical support measures, and those that do not, should be considered; this could also involve the use of machine learning techniques. These analyses could be used to inform practical measures or communication strategies with which to encourage uptake of the scheme. • Predictive analytics combined with practical policy measures, designed to raise scheme awareness and uptake, are likely to be required to maximise the benefits from improved identification. Machine learning can only act as an enabler; if uptake remains low, this could considerably reduce the scope for benefits from the implementation.
Reputational risk	<p><i>Algorithmic tools within a public policy setting are likely to attract public attention, which brings some degree of reputational risk.</i></p> <p>Even if schemes are designed to allocate only benefits to vulnerable populations, if a subset of individuals are unintentionally excluded this can lead to a negative perception or undermine trust in the scheme. The same holds for any perceived biases or opacity within the implementation of the algorithm.</p> <p>For example, New York City received significant political scrutiny when implementing a randomised control trial within the HomeBase programme.⁴⁴ This is despite the objective of the programme being to improve the effectiveness of support for those at risk of homelessness.</p>	<ul style="list-style-type: none"> • Public bodies should be prepared for political scrutiny when implementing machine learning tools, even if this tool seeks only to provide benefits to those in need. • Transparency regarding the usage and limitations of the model is key to mitigate reputational and political risk. If a tool is used directly within the decision-making process for the allocation of support, it is advisable to publish a method statement and the underlying code in order to inform the public on how model outputs are used in practice. This has been the case in Switzerland for refugee allocation purposes.⁴⁵

⁴³ ECO3 Final Stage Impact Assessment. Available at:

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/749638/ECO_3_Final_Stage_IA_Final.pdf

⁴⁴ <https://www.nytimes.com/2010/12/09/nyregion/09placebo.html>

⁴⁵ For example, code to examine the impact of ethnic networks on integration <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/FXVVDQ>

4.3.3 Ethical Challenges

There is a possibility that a machine learning tool could be implemented legally but leads to unethical consequences. Key ethical considerations in terms of potential biases and discrimination, transparency and the level of human oversight have to be made when implementing a predictive tool in a public policy setting. Frameworks such as the UK guide to using AI in the public sector,⁴⁶ the UK Data Ethics Framework⁴⁷ or more practical frameworks such ALGO-CARE that is currently used in policing (Oswald et al. 2018), provide intuitive guidance on the development and implementation of those tools.

Engagement with advisory bodies such as the Centre for Data Ethics and Innovation or the Alan Turing Institute could be further undertaken to facilitate consistency across government departments. There is a strong correlation between ethical and legal considerations, starting with direct references to basic human rights and freedoms, privacy and data handling, setting expectations towards trustworthy artificial intelligence, and through to the European Commission White Paper on Artificial Intelligence.⁴⁸

The following table provides key ethical challenges that need to be considered and draws on the key learnings from the selected case studies, including stakeholder interviews.

Table 5: Potential ethical challenges and mitigations

Challenge	Challenge detail	Recommended mitigations
Conceptual model definition and misspecifications	<p><i>A key principle underlying robustness and safety within the scope of ethical requirements is the principle of prevention of harm.</i></p> <p>This means that a machine learning solution has to be developed with a preventative approach to risks; behave reliably as intended; minimise unintentional and unexpected harm; and prevent unacceptable harm.</p> <p>In seeking to apply machine learning techniques to identify fuel poor households, harm could result from model misspecifications that do not fully reflect patterns in the data and/or data connection to the target variable.</p> <p>As set out in the practical considerations, technical implementation assessments and model validations are required in order to ensure that the solution behaves as intended and to minimise potential harm.</p>	<ul style="list-style-type: none"> At the pre-modelling stage of the project lifecycle, the “conceptual model” should be defined in order to formalise the real-world problem that the machine learning system attempts to fit, with appropriate measures designed to monitor potential “harm” resulting from decisions informed by the model.

⁴⁶ Government Digital Service and Office for AI (2019). Available at: <https://www.gov.uk/government/collections/a-guide-to-using-artificial-intelligence-in-the-public-sector>

⁴⁷ Department for Digital, Culture, Media and Sport (2018). Available at: <https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework>

⁴⁸ European Commission (2020). Available at: https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

Challenge	Challenge detail	Recommended mitigations
Automation based on incomplete information	<p><i>Machine learning models are often limited by data availability and cannot possibly assess all factors that may affect the outcome. Solely relying on algorithmic tools in this instance can lead to unjustifiable actions.</i>⁴⁹</p> <p>For example, the HART tool uses only data that is held by the Durham Constabulary and cannot assess factors such as family circumstances that may influence the behaviour of a person. In making a final decision, custody officers can access additional data sources such as the national police database and may have had previous interactions with the individual. This is just one factor as to why the implication of the tool serves as just one component that informs a human-centric decision-making process, rather than function as the ultimate decision-maker.</p> <p>Across the case studies there was universal consensus that machine learning models should be used to support human decision-making rather than directly acting as the decision-maker, in part as a consequence of this challenge.</p>	<ul style="list-style-type: none"> • Where relevant, it is advised that the implementation of machine learning models should be complemented by human judgement, rather than act as the ultimate decision-maker itself. • A human-centric decision-making framework in this instance could refer to the manual verification of those fuel poor households identified with the highest probability by the machine learning model, prior to offering ECO support. • Alternatively, machine learning models could provide further evidence to inform Local Authority decision-making within the LA Flex mechanism. • There is a larger scope for automation within the WHD Broader Group, given that this represents financial assistance with more strict eligibility criteria. In this case, a suitable challenge mechanism may be sufficient to mitigate this risk (as currently in place for the Winter Fuel Payment scheme). • It should be considered if any other criteria would need to be assessed within the use of the model and determining appropriate actions. For example, what weighting will model outputs hold in determining the next best action, and what other information might be available to policy teams that cannot be accounted for in the modelling.
Presence of false negatives	<p><i>In developing a machine learning tool, it is not practically feasible to produce a model that avoids, to at least some extent, the presence of false negatives.</i></p> <p>A false negative represents the case where a model predicts a household to not be fuel poor, but in practice the household is fuel poor. This could lead to the possibility of households in need being excluded if a tool is used for automated decision-making processes, for example.</p> <p>There are a number of trade-offs to consider in this instance. On one hand, whilst the optimality criteria could be set to minimise the proportion of false negatives (BEIS itself achieved less than 10% false</p>	<ul style="list-style-type: none"> • Before the introduction of a predictive tool in the fuel poverty context, it should be considered what would represent an acceptable threshold for model performance, in particular the trade-off between false negatives and positives. • The threshold for acceptability may depend on how model outcomes are to be used, in particular the degree of automation involved. If a decision to allocate support maintains some degree of human involvement or verification,

⁴⁹ As Oswald et al. (2018) put it “Inconclusive evidence leading to unjustified actions”

Challenge	Challenge detail	Recommended mitigations
	<p>negative predictions within the 2017 study), this will increase the proportion of false positives (in the same study, for the households predicted to be fuel poor, only c.25% actually were). This raises the question as to the tolerance that an organisation has for the presence of false negatives and the mitigations required to address these observations.</p> <p>This challenge was observed across the case studies identified. For example, the HART tool is programmed to minimise false negatives in order to avoid the misclassification of high-risk individuals. The initial model used a ratio of approximately two cautious errors (i.e. the case whereby the model overestimates a risk rating) for each dangerous error (vice versa). However, this by definition leads to an increase in false positives (i.e. an overestimation of those classified as highest risk), which affects the eligibility of people to join the Checkpoint programme.</p> <p>There are two sides to this challenge: on one hand, protecting the public from the highest risk of harm by minimising the most dangerous error could be seen as a priority, but on the other hand, there are clearly ethical considerations to be made when using a tool that deliberately overestimates the risk of individual offenders.</p>	<p>a lower threshold might be tolerated (for example if model outputs are verified prior to the offer of ECO support).</p> <ul style="list-style-type: none"> • If model outputs are to be used directly for decision-making purposes, this would likely require a challenge mechanism to be developed whereby individuals can manually apply for a review of their eligibility or a reopening of their case (for example, if assignment was automated for the WHD Broader Group).
Data biases	<p><i>There is an inherent possibility in any machine learning implementation that underlying datasets could include biases or not be fully representative of the population under consideration.</i></p> <p>Input data quality is a primary concern for building reliable machine learning models because success depends on the input data, the model itself, and the way the model is used. Poor quality input data or data that is not fit for purpose can result in problems ranging from biased to unreliable outcomes. This could be particularly problematic if these outcomes disproportionately affect particular subsets of the population.</p> <p>Data used for training and operation of an artificial intelligence solution may suffer from pre-existing historic biases, incompleteness or limited governance. Moreover, algorithms may pick up existing data patterns and add their own algorithmic biases.</p> <p>Missing observations that are correlated with the underlying characteristics of the population, or whether the dataset is able to capture the full population, are just two examples of possible biases.</p>	<ul style="list-style-type: none"> • In line with the overarching recommendations, the adoption of a formal ethical review process should be considered whereby underlying datasets are subject to a review for potential sources of bias. This would be to ensure that input data is fit for purpose. • Inclusion, non-discrimination and diversity should be enabled throughout the project lifecycle, which relates to the basic fairness principle. Three broad categories of ethical concerns fall into this area: avoidance of unfair bias, accessibility and universal design, and stakeholder participation. • As set out in the 2017 BEIS study, data from the EHS may underweight the fuel poor population or not be fully representative of the households. BEIS should consider how to adapt for this characteristic, for example through reweighting or subsampling.

Challenge	Challenge detail	Recommended mitigations
	<p>These can result in a range of ethical issues, as covered in research by the European Union Agency for Fundamental Rights.⁵⁰</p> <p>In the fuel poverty context, those who are least engaged or aware of their eligibility for support measures are likely to be those that are most in need of assistance; these households may not previously have applied for relevant benefit proxies, ECO or the WHD Broader Group, for example. As a result, there is a possibility that these households may not be fully represented within the underlying data.</p>	<ul style="list-style-type: none"> • Across the case studies identified, underlying datasets were subject to a number of review and validation procedures to identify and subsequently mitigate potential sources of bias.
<p>Algorithmic bias and discrimination</p>	<p><i>Algorithmic bias describes systematic and repeated errors that create unfair outcomes, including privileging one group of users over another. Machine learning models can exacerbate existing biases as they make predictions on historical data, and are updated to incorporate decisions and outcomes determined in previous iterations.</i></p> <p>Biases in historical data, model features and errors (i.e. false negatives and false positives) can potentially lead towards discriminative decisions against particular groups. This bias may be unintentional; even if sociodemographic or protected characteristics (i.e. age, disability, gender reassignment, marriage and civil partnership, race, religion or belief, sex, sexual orientation) that may cause bias are not used in the model, there may be proxies for these characteristics that unintentionally deliver the same effect.</p> <p>There is evidence that instances of fuel poverty may vary across protected characteristics. For example, the 2019 fuel poverty statistics (table 24) indicate that of ethnic minority households, 20% are fuel poor; the corresponding figure is 9.7% for white households.⁵¹ The potential for biases amongst protected characteristics should be treated with caution.</p> <p>Moreover, if particular groups have been disproportionately targeted in the past, the algorithm may incorrectly classify these in future, or skew the results towards targeting similar groups of people again. This can lead to a feedback loop accentuating existing biases as the model is continually updated with new data that incorporates decisions already made by the model, which could increase the risk of discrimination or exclude potentially vulnerable households.</p>	<ul style="list-style-type: none"> • The adoption of a formal framework which includes an independent ethical review of the algorithm and underlying data should be considered. • Model validation assessments should be conducted to identify potential or ongoing biases in the model (i.e. if groups are discriminated against and how often they are missed out). Data should be refreshed on a periodic basis. The results of the assessment can be used to further develop the predictive tool or to formulate certain actions to mitigate biased outcomes. • Ethical and technical reviews can set effective measures for assessing the risks to protected groups and the selection of the “best” model in terms of performance, fairness and transparency. Understanding the ways in which different classification algorithms work can assist in capturing the associated risks of biases. • Ethical guidance such as the UK guide for using AI in public sector or practical frameworks such as ALGO-CARE can be used at the start of the implementation process to document the considerations made in this regard. • In the Portuguese case, the research suggested that parity on the proportion of false negatives and false positives across different groups could be achieved to mitigate the risk of an algorithm discriminating against particular subgroups. Protected characteristics were also kept in the model to allow

⁵⁰ For example: 'A European approach to excellence and trust'. Available at: https://fra.europa.eu/sites/default/files/fra_uploads/fra-2019-data-quality-and-ai_en.pdf

⁵¹ Fuel Poverty Statistics England (2019) – most recent data available from 2017. Available at: <https://www.gov.uk/government/collections/fuel-poverty-statistics>

Challenge	Challenge detail	Recommended mitigations
	<p>For example, Durham Constabulary initially used Experian’s Mosaic datasets, which included a residential postcode. This feature could be viewed as indirectly related to measures of community deprivation. This also risked generating a negative feedback loop. If a police force responds to classifications by targeting resources on the highest risk postcode areas, more individuals from those areas will be subject to attention and be arrested than those living in lower risk, untargeted neighbourhoods.⁵²</p> <p>Furthermore, the assessment of the Portuguese decision-support tool to predict long-term unemployment identified a potential concern with possible systematic biases incorporated into the model. In this case, as particular groups of individuals are more often registered with IEFP (for example, considerably more women than men), and some characteristics of these individuals may have led to longer unemployment in the past, the concern is that this could lead to biases in future predictions against these groups.</p>	<p>for bias auditing and for future comparisons to human-only decisions.</p>
<p>Transparency</p>	<p><i>In many settings, machine learning techniques are able to outperform statistical methods such as logistic regression. However, a common issue of these techniques is referred to as the "Black Box" problem, where the process between input and output is not completely explainable.</i></p> <p>The requirement for transparency represents the principle of AI “explainability” which is widely adopted in the ethical field. This encompasses the transparency of all elements relevant to an AI system: the data, internal workings of the model and application of the modelling outputs. It is a multifaceted requirement that can raise a variety of related risks.</p> <p>Generally, transparency is considered in terms of three building blocks of traceability of data and processes, explainability of processes and decisions, and communication to all stakeholders affected directly and indirectly by the AI solution.</p> <p>There is often a trade-off between explainability and predictive power of a model. For example, statistical regression methods tend to be relatively transparent in producing output but are limited in their predictive capability. On the other hand, complex machine learning techniques such as deep neural networks may be able to deliver</p>	<ul style="list-style-type: none"> • Measures should be undertaken to ensure that the implementation of machine learning is delivered in a transparent manner. This could include published research reports, together with the underlying code and methodology, especially in the case whereby the tool is used for direct decision-making purposes. • Transparency not only applies to the methodology but also the underlying data and variables created. For example, it should be ensured that variables calculated within third party datasets are fully understood (e.g. data on income may be taken from other external sources which could include measurement error). • This principle of transparency in feature development also applies to those calculated by BEIS – any features used should be clearly justifiable using economic rationale. • The need for transparency in a public policy setting may suggest the use of more transparent supervised machine learning methods in this instance. This is, however, an evolving area; not only has considerable research been undertaken to improve the explainability of more complex

⁵² Adapted from Oswald et al. (2018).

Challenge	Challenge detail	Recommended mitigations
	<p>improved predictive power, but may be less explainable in a public policy setting.</p> <p>Across the case studies, government bodies made details of the model publically available to make the process as transparent as possible. For example, the State Secretariat for Migration (SEM) announced the pilot of the machine learning tool on the government website and the research team has published a paper, the code and supplementary material documents that outline details of the methodology and data.</p>	<p>unsupervised techniques, but it is often the case in academic research that parallel models are designed for demonstration purposes to provide a simplified explanation of the method in question.</p>
<p>Automation bias</p>	<p><i>Those using the outcomes of machine learning models on a daily basis may not have a sufficient technical background or understanding of model implications, which could lead to decisions based on incorrect interpretation.</i></p> <p>Machine learning models have the potential to inform and enable consistent decision-making across support schemes. The balance between predictive analytics and professional human judgement is often essential when implementing a machine learning tool in a public policy setting. However, the usage of these models can lead to “Automation bias” in decision-making if a tool provides a ‘recommendation’ that is subsequently fully trusted (and not challenged) by an end-user.</p> <p>For example, if outputs are to be used to inform the offer of ECO support, the factors driving the classification of a household as fuel poor should be understood. This would help to understand any potential sources of error, challenge the outputs and refine the process in future iterations.</p> <p>Across the case studies, communication with case workers and the provision of sufficient training was noted as a possible mitigation to the risk of automation bias. Moreover, how outputs were presented was often found to influence how those using the model reacted to the findings.</p> <p>For example, the initial HART tool uses a traffic light interface that provides case workers with a red, amber or green light for high, moderate or low risk respectively, which may ‘nudge’ an end-user, even if information is also provided as to the use of the tool.</p> <p>The Portuguese IEFPP in partnership with Nova SBE is also continuing research in this area, in particular how a tool should be designed to facilitate engagement with case workers. The study has developed a</p>	<ul style="list-style-type: none"> • Training should be provided to those using the machine learning tool, including the awareness of potential errors and how the model develops outcomes. The objective of this would be to ensure that outputs are used in a consistent and justifiable manner. • Further consideration should be given to the presentation of model outcomes. For example, case study evidence suggested that the degree of uncertainty within a prediction should be presented alongside the classification or probability to those using the tool. • BEIS could also consider following the Portuguese case study example, and provide a dashboard that shows factors contributing to the classification of a household as fuel poor (or not fuel poor) and how these may have changed over time. • In communicating an advisory tool, it is important to be cautious in the use of language; Babuta and Oswald (2020) suggest that statistical forecasting systems based on algorithms should not be described as ‘predictive’ or ‘assessment’ tools, but more accurately as ‘classification and prioritisation systems’, to emphasise that human involvement and challenge remains a vital component of the decision-making framework.

Challenge	Challenge detail	Recommended mitigations
<p>Accountability – ethical perspective</p>	<p>dashboard that displays an overall risk score for an individual, their risk history and individual factors that serve to affect the risk level.</p> <p><i>Accountability of machine learning solutions is a necessary element that complements other requirements and is closely linked to the principle of fairness.</i></p> <p>This means that there should be mechanisms in place to ensure responsibility and accountability for machine learning systems and their outcomes, both before and after their development.</p> <p>Accountability encompasses auditability of algorithms, data and processes, minimisation and reporting of negative impact, ethical trade-offs and redress.</p> <p>Across the case studies, a number of measures were taken to ensure accountability including the publication of underlying model code and the establishment of named process owners.</p>	<ul style="list-style-type: none"> • Auditability requires that all of the inputs, process of turning them into outputs, ultimate outcomes and all associated analysis can be articulated. The public use of the solution also suggests that auditability by the public may be required. • Challenge mechanisms should be in place in order to cover any unjust adverse impacts with accessible means. Redress options should be communicated openly and clearly. This is required to foster public trust in the public service, whether delivered by people or technology.
<p>Societal and environmental wellbeing</p>	<p><i>As AI systems represent "intelligent agents", the broader society and environment can be considered as indirect stakeholders of the project.</i></p> <p>The trustworthy AI approach suggests three building blocks to societal and environmental wellbeing: sustainable and environmentally friendly AI, social impact, and society and democracy.</p> <p>With respect to fuel poverty targeting, given its public nature, the social impact may create many of the requirements and associated risks. According to the UK guide for responsible design and implementation of AI systems in the public sector,⁵³ there are four main values that have to be respected during the implementation of an AI solution:</p> <ul style="list-style-type: none"> • Respect the dignity of individual persons; • Protect the priority of social values, justice, and the interests of the public; • Connect with each other sincerely, openly, and inclusively; and • Care for the wellbeing of each and all. 	<ul style="list-style-type: none"> • One of the implications of these values is that the algorithm should be shown to benefit society either directly or indirectly. • Should it be the case that the algorithm leads to unintentional "non-assignment" (for example, through random or systematic errors), these issues should be recognised and a clear way of remediation should be developed. • The aggregate benefit of the algorithm use should be shown to exceed the cost of it, including the negative impact of its outcomes on people. • The guideline for AI use in the UK public sector⁵⁴ provides a range of recommendations in this regard. It is not legally binding, however, it is suggested that the solution follows typical approaches to reflect the specifics of public service in the adoption of AI.

⁵³ Alan Turing Institute (2019). Available at: https://www.turing.ac.uk/sites/default/files/2019-06/understanding_artificial_intelligence_ethics_and_safety.pdf

⁵⁴ UK Government Digital Service and the Office for Artificial Intelligence (2019). Available at: <https://www.gov.uk/government/collections/a-guide-to-using-artificial-intelligence-in-the-public-sector>

4.3.4 Legal Challenges

There are a number of underlying legal frameworks that should be considered both prior to and throughout the implementation of machine learning methods in a public policy setting, including the European Convention of Human Rights (ECHR), the Equality Act 2010 (together with the Public Sector Equality Duty 2011) and data protection regulations such as the General Data Protection Regulation (GDPR) and the Data Protection Act 2018 (DPA), amongst others. This legislation should be considered together; the DPA often serves to supplement or amend provisions under GDPR, for example. The following table outlines a number of considerations when using predictive tools, based upon publically available sources such as the Information Commissioner’s Office (ICO) and learnings from the selected case studies.

For the avoidance of doubt, this section does not represent legal advice: independent legal advice should be sought on the proposed use of the model and required actions under the applicable legal directives or regulations at the time of implementation. The challenges below do not represent an exhaustive list and the analysis is only applicable at the time of writing; further or amended legislation may be applicable in future periods. The potential mitigations suggested in this section should also be interpreted in this limited manner.

Guidance on the steps that can be taken to align with privacy regulations, for example, is available from the Information Commissioner’s Office (ICO), the UK’s independent authority that upholds information rights in the public interest. For example, the ICO sets out that a Data Protection Impact Assessment⁵⁵ should be conducted for data processing that is likely to result in a high risk to individuals (e.g. automated decision-making, personal data usage, data matching, and usage of data concerning vulnerable data subjects). The ICO notes that *“It is also good practice to do a DPIA for any other major project which requires the processing of personal data.”*

As a general mitigation, legal considerations should be made at the earliest possible stage within a model implementation, both to mitigate the challenges outlined below and identify any additional barriers. Otherwise, models may require adjustment or redevelopment at a late stage in the process, which may either introduce infeasibilities or increase the necessary time and resources.

Table 6: Potential legal challenges and mitigations

Challenge	Challenge detail	Recommended mitigation
Alignment with European Convention on Human Rights	Human rights legislation is becoming increasingly important in a machine learning context – these principles often lay the foundation for the consideration of legal and ethical issues, with particular focus on ECHR Article 8 - Right to respect for private and family life: ⁵⁶ <i>“There shall be no interference by a public authority with the exercise of this right except such as is in accordance with the law and is necessary in a democratic society in the interests of national security, public safety</i>	<ul style="list-style-type: none"> BEIS should consider these “fair balance” principles, namely the trade-off between the objectives of the tool and privacy considerations – this should be included within a formal ethical and legal review process. Legal advice should be sought regarding the alignment with ECHR prior to implementation.

⁵⁵ ICO Guidance: <https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/accountability-and-governance/data-protection-impact-assessments/>

⁵⁶ ECHR Article 8. Available at: https://www.echr.coe.int/Documents/Guide_Art_8_ENG.pdf

Challenge	Challenge detail	Recommended mitigation
	<p><i>or the economic wellbeing of the country, for the prevention of disorder or crime, for the protection of health or morals, or for the protection of the rights and freedoms of others.”</i></p> <p>In a recent case in the Netherlands,⁵⁷ the Dutch court decided to order the immediate halt of a welfare fraud surveillance system based on ECHR Article 8 violations, following concerns regarding a lack of transparency and that existing legislation contained insufficient safeguards against privacy intrusions. In particular, the system did not pass the test required by the ECHR of a “fair balance” between its objectives, namely to prevent and combat fraud in the interest of economic wellbeing, and the violation of privacy that its use entailed.</p> <p>Even for policies whereby tools are used with the intention for only positive consequences, for example the provision of benefits, these still have the potential to violate ECHR Article 8 should a tool make opaque classifications based upon social status, for example, which individuals may consider damaging or intrusive.</p>	<ul style="list-style-type: none"> • As set out in the UK guide for using AI in the public sector, BEIS should consider publishing elements of the model and supporting documentation to ensure transparency. This was recently the case in Switzerland for the case of optimising refugee allocation.
Equality Act	<p><i>Under the Equality Act 2010 it is illegal to discriminate against individuals with protected characteristics (i.e. age, disability, gender reassignment, marriage and civil partnership, race, religion or belief, sex, sexual orientation).⁵⁸</i></p> <p>The Public Sector Equality Duty also requires public bodies to have due regard to the need to eliminate discrimination, advance equality of opportunity and foster good relations between different people when carrying out their activities.⁵⁹</p> <p>A potential risk of algorithms is that they could incorporate data or proxies that are strongly correlated with these attributes, which can lead to such biases. This may be a particular issue if false negatives or positives are subsequently more present across specific subgroups. As set out in the previous section, there is evidence that instances of fuel poverty are likely to vary across these characteristics; this should therefore be treated with caution.</p> <p>The HART tool experienced this issue when employing Experian’s Mosaic data in the model, which included behaviour indicators together with residential postcodes that categorised groups of people based on their</p>	<ul style="list-style-type: none"> • BEIS should consider performing an assessment of potential biases within the ongoing testing and validation of the model. • In the case of long-term unemployment in Portugal, it was recommended that protected features could be used in order to test and measure the extent of potential bias within the model and define mitigations accordingly. • Particular attention should be paid to third party datasets – for example, the methodology by which any features used from third parties have been calculated should be fully understood, and whether these variables could be correlated with individuals’ protected characteristics.

⁵⁷ <https://www.theguardian.com/technology/2020/feb/05/welfare-surveillance-system-violates-human-rights-dutch-court-rules>

⁵⁸ Equality Act 2010. Available at: http://www.legislation.gov.uk/ukpga/2010/15/pdfs/ukpga_20100015_en.pdf

⁵⁹ Public Sector Equality Duty 2011. Available at: <https://www.gov.uk/government/publications/public-sector-equality-duty>

Challenge	Challenge detail	Recommended mitigation
	<p>location; this could be interpreted as a proxy for community deprivation. The Durham Constabulary decided to limit the use of the postcode feature after a model validation exercise.</p>	
<p>Privacy and personal data: Automation</p>	<p><i>Both GDPR and the DPA address the issue of decisions based on automated processing. This is just one example where legislation should be considered together.</i></p> <p>GDPR Article 22(1)⁶⁰ on “Automated individual decision-making, including profiling” sets out that “data subjects have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her”. This would not apply however if the decision, amongst other exemptions, is based on the subject’s explicit consent.</p> <p>The DPA Section 14 also makes provisions for the purposes of Article 22, in particular exception from Article 22(1) of the GDPR for significant decisions based solely on automated processing that are authorised by law and subject to safeguards for the data subject’s rights, freedoms and legitimate interests.⁶¹</p>	<ul style="list-style-type: none"> • BEIS should consider measures to ensure that any automated model is in compliance with GDPR article 22, unless subject to exemptions. As set out in the directive, this would likely include the development of a suitable challenge mechanism whereby non-recipients can challenge the non-receipt of support based on a decision made by the algorithm, or may require the explicit consent of the data subject. For example, BEIS should consider the implications of this legislation should automation be pursued within the WHD Broader Group.
<p>Privacy and personal data: Right to inform / object</p>	<p><i>Articles 12 – 23 under GDPR set out the rights of the data subject with regards to data processing and automated decision-making.⁶² In particular, these include provisions for the right of data subjects to be informed or to object to the processing of personal data in a number of circumstances.</i></p> <p>For example, GDPR Articles 13 and 14 set out transparency obligations for the data controller in relation to the processing and storage of personal data. In particular, data subjects have the right to be provided with information on the purpose or nature of personal data processing and the identity and contact details of the data controller. GDPR Article 16 also includes a right for individuals to have inaccurate personal data rectified, or completed if it is incomplete.</p>	<ul style="list-style-type: none"> • Where required, procedures should be in place to ensure that all transparency obligations and data subject rights are met, based on legal advice. • In line with standard data processing requirements, BEIS should be prepared for requests from individuals affected as to any personal data held about them and how this data is used in practice. This should be part of the overall governance framework and the data governance process. • The exact requirements are likely to depend on the final usage of the machine learning algorithm, and further legal advice would be required to determine necessary actions. • Within the selected case studies, these requirements did not represent a prohibitive barrier to the planned operation of machine learning tools, as long as data processing was

⁶⁰ GDPR articles available at: <https://gdpr-info.eu/>

⁶¹ UK Data Protection Act (2018). Available at: <http://www.legislation.gov.uk/ukpga/2018/12/section/14/enacted>

⁶² GDPR Article 12. Available at: <https://gdpr-info.eu/art-12-gdpr/>

Challenge	Challenge detail	Recommended mitigation
	<p>Moreover, GDPR Article 21 states that the data subject has the right to object at any time to the processing of personal data concerning him or her, including profiling based on those provisions.</p> <p>Whether this right applies to the particular case in question however depends on the purposes and lawful basis for processing. For example, where processing personal data is “for scientific or historical research, or statistical purposes the right to object is more restricted.”⁶³</p> <p>Across the case studies identified, these provisions (or similar) were not prohibitive for the operation of the tools employed, as long as data processing, for example, was performed in line with necessary guidelines.</p>	<p>compliant with the legislation and appropriate mechanisms were in place.</p>
<p>Privacy and personal data: Data protection by design and default</p>	<p><i>Data protection by ‘design and default’ is a legal requirement under GDPR (i.e. data protection has to be embedded from design throughout to application of any use of personal data).</i>⁶⁴</p> <p>This means that appropriate technical and organisational measures should be put in place to implement the data protection principles and safeguard individual rights.</p> <p>There are a number of measures by which data safeguards can be put in place, such as anonymisation or pseudonymisation. As set out in GDPR Recital 26.⁶⁵</p> <p><i>“The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable.”</i></p> <p>However, although pseudonymisation can enable greater flexibility in data use, this may not fully remove the risk that a person cannot be identified by other means. As set out in Recital 26:</p>	<ul style="list-style-type: none"> • In line with ICO guidance, a Data Protection Impact Assessment that assesses proportionality, compliance measures and identifies measures to mitigate risk is required for data processing that is likely to result in high risk to individuals.⁶⁶ The ICO recommends a DPIA for “any other major project which requires the processing of personal data.”⁶⁷ • Stakeholders that will have access to model outcomes should be defined to ensure that personal data generated by a model is treated in the same manner as personal data collected through regular channels. • It is recommended that legal input is integrated into the development and design of the tool, as opposed to an ex-post assessment. • Anonymisation or pseudonymisation of personal data may enable more flexible use of data, however pseudonymisation may not be sufficient to remove the risk that a data subject represents an identifiable natural person.

⁶³ ICO Guidance: <https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/right-to-object/>

⁶⁴ <https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/accountability-and-governance/data-protection-by-design-and-default/>

⁶⁵ <https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/what-is-personal-data/what-is-personal-data/>

⁶⁶ Criteria which may act as indicators of high risk processing are for example, automated decision-making, personal data usage, data matching, and usage of data concerning vulnerable data subjects or applying new technologies such as machine learning.

⁶⁷ <https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/accountability-and-governance/data-protection-impact-assessments/>

Challenge	Challenge detail	Recommended mitigation
	<p><i>"Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person"</i></p> <p>It should also be recognised that model outcomes may generate new personal information (e.g. data that refers to a decision made about an identifiable individual). Model outcomes that are passed on outside of the immediate model user or owner group may lead to unintended negative consequences.</p>	
<p>Privacy and personal data:</p> <p>Personal data minimisation and purpose limitation principles</p>	<p><i>The GDPR data minimisation principle sets out that data owners must limit personal data collection, storage, and usage to data that is relevant, adequate, and absolutely necessary for carrying out the purpose for which the data is processed.</i></p> <p>Further, the GDPR purpose limitation principle (GDPR Art.5(1)(b)) sets out that personal data shall be collected for specific, explicit and legitimate purposes and not further processed (for achieving purpose in the public interest, scientific or historical research or statistical purpose) in a manner that is incompatible with this purpose.⁶⁸ Data sharing and matching can be undertaken when it complies with all underlying data protection regulations in this regard.</p> <p>Public bodies across the case studies often limited data use to datasets already held by the department or where data-sharing agreements were already in place. It is noted however that data sharing often represents a barrier due to practical issues (e.g. cultural or data quality) as opposed to legal issues.</p>	<ul style="list-style-type: none"> • BEIS should consider whether the collection and storage of personal data is proportional to the purpose of the model, in line with GDPR. This would include assessing the impact of additional data on model accuracy and periodically reviewing the data held. • Should data sharing not be feasible given further legal or other considerations, BEIS should determine whether it is possible to build a machine learning model that meets their accuracy thresholds using publically available data or datasets already held by the Department, together with any implications for the implementation design.
<p>Privacy and personal data:</p> <p>Accountability</p>	<p><i>Organisations are required to take responsibility and accountability for the usage of personal data.</i></p> <p>The GDPR Accountability principle sets out that an organisation needs to be able to demonstrate compliance with data protection principles. Specific guidance on this issue is also available from the ICO, including a checklist of appropriate steps to demonstrate compliance.⁶⁹</p>	<ul style="list-style-type: none"> • Named owners should be developed for each stage of the implementation process, particularly with regards to the usage of personal data and the outcomes of the machine learning tool. It is recommended that a "senior person" be appointed that oversees the rollout of the scheme and coordinates the usage of the tool and underlying data across internal stakeholders.

⁶⁸ GDPR Article 5. Available at: <https://www.privacy-regulation.eu/en/article-5-principles-relating-to-processing-of-personal-data-GDPR.htm>

⁶⁹ ICO Guidance: <https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/accountability-and-governance/>

Challenge	Challenge detail	Recommended mitigation
	<p>Across the case studies a “senior person” was named within the governance structure who was responsible in overseeing the rollout of the predictive tools, for example.</p>	
<p>Ongoing developments in the legal framework</p>	<p><i>Legal frameworks in relation to the use of machine learning are continually evolving. The use of these techniques may be subject to more direct regulation in future.</i></p> <p>Whilst this is a gradual process, institutions seeking to implement these techniques may wish to consider wider engagement with government, advisory and regulatory bodies to determine the potential impacts of changes to the regulatory environment.</p>	<ul style="list-style-type: none"> • BEIS should consider engagement with CDEI and DCMS with regards to the ongoing development of potential regulations and the impact that these may have on the use of machine learning for their purposes.

5 Scope for benefits

The improved targeting of fuel poor households can not only improve the provision of support measures but also achieve wider benefits to society. This section analyses the scope for benefits from machine learning implementation and provides a worked example to quantify the potential impacts.

5.1 Overview of potential benefits

There are widespread potential benefits from applying machine learning to better identify fuel poor households, over and above the general benefits described in Section 2.

Firstly, machine learning techniques can lead directly to the better identification of fuel poor households and hence the improved allocation of targeted support. This would be of particular relevance for the ECO and WHD schemes.

In the case of ECO, obligated suppliers incur significant search costs in identifying eligible properties for the delivery of energy efficiency installations. Furthermore, customers who are not eligible for the WHD Core Group are required to apply directly to their energy supplier to receive support within the Broader Group. Both of these policies could therefore stand to benefit should machine learning better identify fuel poor households; not only could this facilitate the improved provision of support but also potential reductions in search cost. Within future iterations of the schemes, the insights from machine learning could further be used to inform the scheme eligibility criteria to improve targeting efficiency.

Improvements in the support offered to the fuel poor can, in turn, provide benefits for these households including increased consumer welfare (for example through being more able to heat one's home), together with energy savings that result in lower bills (through the installation of energy efficiency measures) under the ECO scheme. More widely, these household benefits can further contribute towards wider government objectives such as the 2050 net-zero greenhouse gas target⁷⁰ and increased health and social wellbeing.

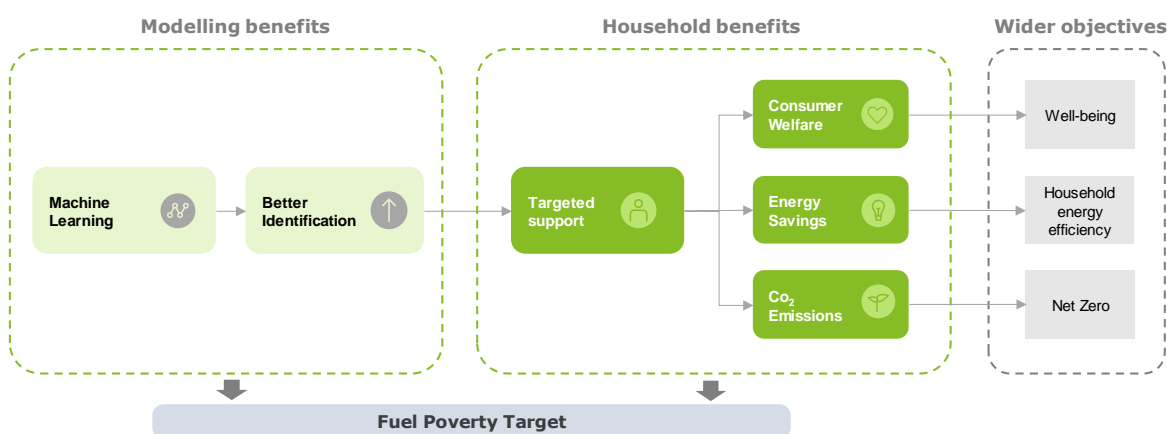
Importantly, given the typically larger energy needs of fuel poor households, net energy savings and emissions reductions could likely be achieved not only by increasing the number of households receiving support under ECO, but also by changing the composition of the scheme to include a greater proportion of fuel poor households, should budget constraints remain fixed at current levels.

Figure 3 summarises this benefits transmission mechanism; this is of particular relevance for schemes such as ECO which involve the installation of energy efficiency measures.⁷¹

⁷⁰ <https://www.gov.uk/government/news/uk-becomes-first-major-economy-to-pass-net-zero-emissions-law>

⁷¹ It is noted that this transmission mechanism may not apply to all current fuel poverty schemes. For example, the Winter Fuel Payment scheme provides bill support for pensioners and can increase welfare, but is not likely to lead to energy or carbon emission savings.

Figure 3: Benefits transmission mechanism to fuel poor households



Source: Deloitte analysis

An assumption-based example that sets out an approach to illustrate the potential scope for these benefits is presented in Appendix 4. This example illustrates how improved identification under machine learning could change the composition of a hypothetical future iteration of the ECO scheme towards the fuel poor, which in turn would deliver both bill savings to this group and a net benefit to society, within existing budget constraints. The example also illustrates what the equivalent search costs to suppliers would have been to identify a similar number of fuel poor households.

This example should not be interpreted as indicative of the scale of benefits that could be expected in practice; further work is required to develop a formal cost-benefit analysis that considers the implementation of machine learning techniques. Appendix 4 sets out full details of the methodology and assumptions employed to estimate these figures.

Assuming that the implementation of machine learning results in 25% of fuel poor households that are currently ineligible for ECO3 becoming eligible (and a corresponding number of non-fuel poor, currently eligible households becoming ineligible to reflect fixed budget constraints), then this illustrative example would result in:

- An additional c.335,000 fuel poor households being identified for support, of which c.100,000 accept the offer of energy efficiency installations.
- Total annual bill savings of c.£20m across these households (£200 per household), with a net annual impact (societal benefit) of £9.7m (£97 per household) taking into account the reduced eligibility of non-fuel poor households and their relative incomes.
- A net present value of c.£143m if net annual impacts are considered over a 20-year horizon.
- Potentially large scope for reductions in search cost; under existing policies, it is estimated that identifying the c.335,000 additional fuel poor households referenced above would cost c.£100m, indicating scope for cost efficiencies.

5.1.1 Modelling benefits

There are a number of both quantitative and qualitative benefits that could result directly from the implementation of machine learning techniques. These include:

i) Improved identification:

A challenge to the operation of some existing fuel poverty schemes is that households may not be aware of their eligibility for support. Any improvement in the identification of fuel poor households could therefore greatly improve the provision of these support measures, particularly for ECO and the WHD Broader Group, whereby suppliers either need to search for eligible customers (ECO) or

customer are required to apply for support (ECO and WHD Broader Group). Machine learning techniques can improve identification by better representing the relationship between outcomes and underlying inputs.

Across the case studies, the implementation of machine learning techniques has in many cases led to improved outcomes, over and above how the policy was previously operated. In the highest case it was found that the implementation of machine learning to refugee assignment in Switzerland could improve (employment) outcomes by up to 73% compared to existing methods:

- **United Kingdom:** The HART tool facilitates the identification of eligible people for the Checkpoint programme in Durham, by informing an officer's decision with a risk classification. The overall validated accuracy of the initial model was 62.8% (Oswald et al. 2018), however, the tool was particularly effective at avoiding false negatives (i.e. only c.2% of offenders predicted to be at low risk subsequently displayed high-risk behaviour). In the policing context, minimising this most 'dangerous' form of error was considered a priority to protect the public from the risk of harm. The key benefit of the HART tool has been to improve the evidence base provided to custody officers to inform their decision-making.
- **Portugal:** The decision-support tool seeks to identify people at highest risk of becoming long-term unemployed. Supervised methods (gradient boosted trees) delivered the highest predictive power in this instance, over and above standard techniques such as logistic regression. Early evidence suggests that the use of the tool to inform case workers' judgement could improve the identification of those at highest risk of becoming long-term unemployed, over existing policy measures (the extent to which is subject to ongoing research and development). This would help IEFP to allocate resources more effectively and provide tailored services to those at highest risk.
- **Switzerland:** Model evaluations of the refugee resettlement tool used by the Swiss State Secretariat for Migration have shown that the predictive tool could increase refugees' long-term (third year) employment rate by up to 73% compared to the previous system that was based on random assignment. A randomised control trial is currently ongoing to test the algorithm in practice. The Swiss government is likely to implement the predictive tool if the trial is successful in showing the anticipated benefits to refugees and wider society.
- **United States:** The New York City Department of Homeless Services developed a statistical scoring system to allocate support to families at highest risk of homelessness.⁷² This was developed using empirical research (for example, Greer et al. 2016) that used regression techniques to predict shelter entry over a 2-8 year period; these were shown to be as least as effective as worker judgements, increasing correct predictions by 77% and reducing unidentified cases of subsequent homelessness by 85%. The New York City Human Resources Administration is seeking to implement machine learning models to improve the outreach of the scheme in future.

In all of these cases, machine learning methods continue to be tested, trialled or have already been implemented, recognising the scope for improved outcomes over and above existing policy measures.

ii) Fuel poverty target:

A machine learning model that improves the identification of fuel poor households can improve their energy efficiency through the more effective delivery of ECO measures such as loft or cavity wall insulation. This would contribute towards as many fuel poor households as reasonably practicable achieving a minimum energy efficiency rating of Band E by 2020, Band D by 2025 and Band C by 2030 (the "Fuel Poverty Target").

Recent EPC data suggests that the UK as a whole faces a particular challenge to retrofit homes to meet energy efficiency targets, with the Government stating that measures are required "much

⁷² Responsibility for the scheme is now held by the Human Resources Administration (HRA).

further and faster" to improve household energy efficiency.⁷³ Innovative solutions are likely to be required to further deliver progress against the target. Machine learning can assist by improving the delivery of targeted support and through providing an understanding both of where these households are and the characteristics that affect, in practice, whether a household is fuel poor.

As set out in Section 4, a potential challenge to this regards the uptake of schemes such as ECO; households are required to agree to the installation of measures at the property. Evidence from the ECO3 impact assessment suggests that 'findability' rates, namely the proportion of technical potential that can be identified and subsequently installed by the supply chain within any given year, range from c.11% for solid wall insulation up to 100% for central heating measures. Machine learning can only act as an enabler by identifying households for support; the extent to which progress against the target can be improved depends on households accepting the support on offer.

iii) Cost efficiency:

Predictive models can reduce the time and cost required to identify fuel poor households, and lead to the more effective use of existing resources.

This is particularly with regards to supplier search costs incurred when identifying households for ECO support. Within the ECO3 impact assessment,⁷⁴ assumptions in the medium scenario indicate possible average search costs of around £300-350 per "lead" across the measures considered, with the highest case using assumptions as high as £1,000 (in the case of boiler replacements off the gas grid). Whilst identification under machine learning would not be cost-free in practice, this indicates a clear scope for cost efficiencies, which could in future be passed on to consumers.

In the case of WHD, improved identification of fuel poor households eligible for the Broader Group could subsequently lead to greater participation (potentially through automation) or awareness of their eligibility for support schemes. The extent to which this is achieved would be dependent on how machine learning insights are translated into practical measures that influence the participation of this group.

The extent to which cost efficiencies can be achieved would also be driven by the predictive power of machine learning techniques. The inherent presence of false negatives and positives is likely to require that some degree of manual review, or suitable challenge mechanism, be implemented to ensure that policies are providing support to those households that need it the most. This issue is further discussed in section 5.2.

iv) Design of future fuel poverty schemes:

The implementation of a machine learning tool to identify fuel poor households can help to tailor the design of future fuel poor household energy efficiency schemes.

For example, machine learning techniques can facilitate a better understanding of the factors that are most influential in determining whether a household is in practice fuel poor. This could inform the development of adapted eligibility criteria to improve targeting efficiency within future scheme iterations or indeed inform the development of new initiatives.

v) Machine learning framework:

In dedicating resources towards the implementation of machine learning, BEIS would have developed a framework for implementation that can be transferred across other use cases. Governance structures, data storage and modelling resources, for example, could all be repurposed and tailored

⁷³ <https://www.bbc.co.uk/news/uk-50573338>

⁷⁴ ECO3 Final Impact Assessment. Available at:

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/749638/ECO3_Final_Stage_IA_Final.pdf

for future machine learning implementations. Learnings from the practical, ethical and legal challenges and how these were mitigated would also serve to inform future research.

vi) Use of machine learning in public policy:

Ultimately, a successful implementation would help to accelerate further work on advanced statistics and machine learning methods both within BEIS and potentially across other government departments. BEIS could represent a leading example in implementing machine learning best practice to provide support to those in need and provide further evidence for the use of machine learning techniques in the public sector.

5.1.2 Wider social and environmental benefits

vii) Social benefits:

Improved identification and better targeting of fuel poor households can help to provide financial assistance or energy efficiency measures to those in need. Keeping households warm can have a positive effect on wellbeing and health of those households who could otherwise not afford to pay their energy bills. Inadequately heated accommodation can also lead to physical and mental health effects for residents.

BEIS has previously estimated the health impacts through ECO3 energy efficiency measures at around £177m over the course of the scheme,⁷⁵ with the installation of cavity and loft insulation comprising the majority of these benefits. Improved tailored support can further help to improve health outcomes and also indirectly reduce costs to the health care system.

viii) Environmental benefits:

The installation of energy efficiency measures such as loft or cavity wall insulation under the ECO scheme can help to reduce household energy consumption, carbon emissions and contribute to wider government objectives such as the net-zero target, which requires the UK to bring all greenhouse gas emissions to net-zero by 2050.⁷⁶

Since the implementation of ECO at the end of 2012, c.2.8 million measures have been installed in around 2.1 million properties (which include both fuel poor and not fuel poor households), including 92,965 households in 2019. Households with ECO measures often have more than one measure applied (on average 1.3 measures per household receiving ECO support), driving further improvements in efficiency where households take up support schemes.

Better targeting of fuel poor households eligible for ECO through a machine learning model has the potential to increase those numbers further, which in turn would contribute towards wider climate change actions and the achievement of net zero in the UK by 2050.

Not only can identifying further fuel poor households for support drive these environmental benefits, but a net benefit can also be achieved by changing the composition of those receiving support towards the fuel poor. In particular, fuel poor households are typically larger in size (inhabitants) and often include dependent family members, driving a higher energy need relative to other households currently eligible for ECO. As a result, in the case whereby scheme budget constraints remain fixed (and as such the total number of households receiving support remains virtually constant), improving the proportion of fuel poor households receiving support within the scheme would likely drive a net environmental benefit, even if this meant reduced eligibility or participation of non-fuel poor households.

⁷⁵ ECO3 Final Impact Assessment. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/749638/ECO_3_Final_Stage_IA_Final.pdf

⁷⁶ <https://www.gov.uk/government/news/uk-becomes-first-major-economy-to-pass-net-zero-emissions-law>

5.2 Overview of potential trade-offs

Whilst there is a wide potential scope for benefits, the introduction of machine learning methods also involves a number of potential trade-offs.

i) Inclusiveness vs. misidentification

To reduce the possibility of “excluding” eligible recipients, optimisation of the machine learning algorithm could seek to minimise false negatives (for example, as undertaken within the HART policing tool). However, the resulting trade-off is that by seeking to minimise false negatives within the training set, the number of false positive observations (i.e. classifications whereby a household is not fuel poor, but is classified by the model as fuel poor) will likely increase accordingly.

This phenomenon was perhaps exhibited within a previous BEIS (2017) study that sought to classify fuel poor households through machine learning techniques.⁷⁷ The study was able to achieve a relatively low proportion of false negatives (of those households that were actually fuel poor, the model classified only c.10% as not fuel poor), however for every household predicted to be fuel poor, only approximately a quarter actually were, reflecting a large degree of false positive predictions. This phenomenon reflects a model precision rate of only 23.5% and recall of 89.9%.⁷⁸

This trade-off could have wider implications for the operation of fuel poverty support schemes – whilst seeking to minimise false negatives reduces the likelihood of excluding eligible households, the potential resulting increase in false positives could lead to an increase in the cost of delivering the schemes. For example, this may limit the scope for search cost efficiency or impose additional scheme costs through a requirement for the manual review of those identified as fuel poor (and indeed non-fuel poor) by a machine learning tool. As such, this is not just a modelling trade-off but also represents a consideration given existing budget constraints.

A next step for BEIS is to consider the prioritisation and thresholds for false negatives (and indeed positives) that would represent an “acceptable” level of model performance, taking into account these wider policy impacts. A suitable challenge mechanism may also need to be designed whereby non-recipients can reopen an assessment of their individual circumstances.

ii) Accuracy vs. interpretability

Different supervised or unsupervised machine learning techniques could be used for the purpose of identification. In many settings, techniques such as neural networks can offer improved predictive power over and above traditional methods. However, there is often a trade-off between predictive power and the interpretability of these techniques. Unlike common statistical techniques such as logistic regression, there are in many cases no “coefficients” or equivalent with which to transparently assess the marginal effects of individual factors.

It may also be challenging to develop an audit trail as to how an unsupervised learning method, for example, has arrived at a particular classification. In the fuel poverty context, it is therefore pertinent to consider the prioritisation between transparency and classification accuracy, within the methodology selection.

An illustration of a generalised interpretability and accuracy trade-off is shown in Figure 4 for a variety of techniques. This ordering may not always hold true in practice; for example, methods such as logistic regression may outperform machine learning techniques in some instances. The scope for improvements in accuracy from machine learning will likely be related to the quality and quantity of

⁷⁷ BEIS (2017). Available at:

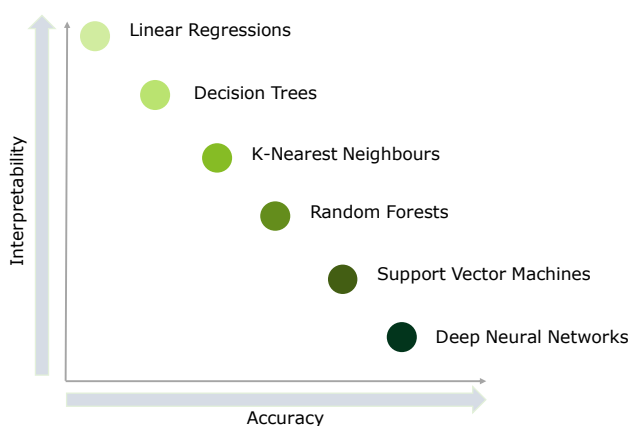
https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/633228/need-framework-annex-a-fuel-poverty-targeting.pdf

⁷⁸ Precision reflects the formula True positives / (True positives + False positives), whereas Recall reflects True positives / (True positives + False negatives).

data used to train the model; these techniques are often able to better capture complex relationships between the output variable and underlying features.

Moreover, it is not always the case that the results of unsupervised models cannot be explained or shown transparently; academic papers often produce models for demonstration purposes in parallel in order to improve the interpretability of results.

Figure 4: Illustration of the Accuracy vs. Interpretability trade-off



Source: Deloitte analysis

iii) Privacy vs. accuracy

From a data science perspective, the more data that is used to train the machine learning model, the more likely it is to capture any statistically significant relationships between the underlying features.

However, collecting additional personal data can have adverse impacts on the privacy of individuals. This is particularly relevant in light of the GDPR data minimisation principle, which sets out that any processed personal data should be adequate (sufficient to fulfil the stated purpose), relevant (has a rational link to that purpose) and limited to what is necessary.

In the fuel poverty context, the right balance between including personal data that would enable improved identification of fuel poor households and excluding data that is not necessary for the purpose of the model should be considered in line with these directives.

iv) Accuracy vs. fairness

Including sociodemographic or personal information within the machine learning model can lead to biased or discriminatory outcomes. Whilst the exclusion of variables that may act as proxies for protected characteristics may help to mitigate this outcome, this may also result in a less accurate model.

The reverse may also be true; if an organisation does not have sufficient information on a minority group, or in this case, those that are eligible but do not currently receive support (e.g. those that may be eligible but have not applied (or been identified) for ECO or the WHD Broader Group), then both fairness and accuracy could be improved by collecting more relevant data.

In developing the machine learning technique and assessing appropriate features, consideration should be made with regards to the impact of including sociodemographic indicators or proxies within the modelling, together with the possibility that certain subpopulations could be underrepresented in the model dataset.

6 Conclusions and next steps

The challenges, mitigations and overarching recommendations identified in this report will inform the next steps within the prospective machine learning implementation.

This report has identified a number of challenges that are likely to be encountered within a machine learning implementation and how these might be mitigated, together with a number of overarching recommendations to inform future phases of work.

Drawing on the learnings from wider research and selected case studies, this report has found:

- **There are a number of relevant use cases whereby machine learning, automation or statistical methods have been applied successfully in a public and social policy context**, often in cases where the final decision has potentially material consequences for those affected. This study has outlined just four detailed examples across policing, long-term unemployment, refugee integration and homelessness prevention.
- **There exists a large potential scope for benefits from the implementation of machine learning techniques**, over and above existing policy measures. This report provides an assumption-based example to illustrate how both the fuel poor and society more widely can benefit from improvements in identification, together with a large potential scope for search cost efficiencies. Implementation could not only inform the design of future fuel poverty schemes but also act as a case study for the use of machine learning in other public sector settings.
- **Although there exists a wide range of potential challenges, evidence does not suggest that any one given challenge represents a barrier that cannot be mitigated to at least some extent within a possible implementation**, assuming the model passes the thresholds for acceptability in terms of predictive power. Learnings from wider research and the four cases studies provide a number of possible mitigations to address the practical, ethical and legal challenges identified in this report.
- **Amongst the challenges identified, ethical challenges regarding potential data and algorithmic biases represent some of the largest risks to implementation**. It should be ensured that a robust framework is in place to monitor the ongoing performance of the algorithm and mitigate potential discriminatory outcomes, together with biases in the underlying data and possible negative feedback loops.
- **Caution should also be applied in seeking to deliver fully automated solutions through machine learning**. This is broadly based on two considerations: that these techniques may not be able to capture the full range of features that determine the target variable, together with the inherent presence of false negative predictions that may lead to the exclusion of particular subgroups. At a minimum, a challenge mechanism should enable non-recipients to re-open their individual cases for review, for example if pursuing automation of the Warm Home Discount Broader Group.

These findings, together with the overarching recommendations set out in this report, inform a number of potential next steps.

- BEIS should continue to test the development of machine learning techniques and consider using a wider range of data sources to improve predictive power. Performance thresholds for

model acceptability will subsequently need to be determined. This is both a statistical assessment but should also consider the desired policy objectives implications for the cost of delivering the schemes.

- Further consideration should be made towards the machine learning implementation design and how outputs will be used to inform policy decisions, including the degree of automation and human oversight. Legal advice should be sought at the earliest possible stage within this process, for example to assess the alignment with GDPR and Human Rights Legislation and determine appropriate actions to ensure that the rights of data subjects are upheld. This would include the development of a formal challenge mechanism where non-recipients could re-open an assessment of their individual cases.
- A formal governance framework should be developed including the division of ownership roles and responsibilities across all stages of the implementation. Development of this framework could draw upon the learnings of established models such as ALGO-CARE, together with engagement with advisory bodies such as the CDEI.

This study has also identified a number of potential areas for possible further research, including:

- *Further case study research:* a number of the case studies identified in Phase 1 of the research were either in development or undergoing initial testing. It may be informative to update this research in future periods once the full scope for benefits has been determined.

Alternatively, research could develop further case studies; this could include examples from other contexts such as the use of machine learning in healthcare, or be selected as a close comparator to the implementation design in this instance.
- *Framework for cost-benefit analysis:* This report has set out an assumption-based example to illustrate the potential scope for benefits from machine learning. In developing a business case for the use of these techniques, this framework could be expanded to develop a formal cost-benefit analysis of the proposed implementation, taking into account the implementation design, the resources required and the cost of delivery.
- *Uptake of fuel poverty schemes:* If the rate of acceptance for support remains low, this can serve to significantly reduce the scope for benefits from improved identification. This is a particular concern for schemes such as ECO which require physical changes to a property. Further research should be considered, potentially including machine learning techniques, to examine the characteristics influencing the rate of uptake.

These exercises could subsequently inform a future impact assessment that would need to be developed for government stakeholders, should the use of machine learning to identify fuel poor households be pursued.

Appendix 1 – Machine learning terminology

A1.1 Defining key concepts

In general terms, artificial intelligence can be defined as the theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making and translation between languages.⁷⁹ This study is particularly concerned with a subset of these techniques, namely machine learning methods for decision-making purposes. In this case, the target variable represents the classification of a household as fuel poor.

Machine learning is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.⁸⁰ The user defines the input data and the target variable; the computer system develops an algorithm that links them. In certain settings, these techniques have been shown to offer improved classification and prediction above traditional methods such as regression analysis.⁸¹

There are three predominant types of 'learning' within this context: supervised, unsupervised and reinforcement learning:

- **Supervised learning:** A supervised learning algorithm takes a known input dataset and its responses to an output variable to learn the classification model that relates them.⁸² In other words, the algorithm 'observes some example input-output pairs and learns a function that maps from input to output' (Russell and Norvig 2016). Examples of these methods include decision trees and random forests, amongst others.⁸³
- **Unsupervised learning:** Unsupervised learning is the opposite of supervised learning, whereby the computer system learns patterns in the input even though no explicit feedback is supplied. These methods, which include deep neural networks, clustering and principal components analyses, are typically more computationally intensive than supervised learning and may deliver more opaque outputs relative to supervised methods.
- **Reinforcement learning:** Reinforcement learning is a cyclical process whereby an agent is presented with an input describing the current state, responds with an action and receives some reward as an indication of the value of its action. The goal of the agent is to maximise the rewards it receives, through this trial and error experience using feedback from its own actions. The use of these techniques to develop recommendation engines has been well publicised in one-off game settings such as Chess and Go.⁸⁴

If data is labelled, namely has been tagged with identifying characteristics, properties or classifications (for example, a variable indicating whether a person is employed), supervised learning methods may be most appropriate. If data is unlabelled and the objective is to identify a structure (for example, a set of uncaptioned photos where the objective is for the algorithm to determine whether the image contains X, Y or Z), this could indicate the use of unsupervised methods. Reinforcement learning is most applicable in situations where the objective is to gather information from repeated interactions within an environment, as in the Chess example above.

⁷⁹ Source: Oxford English Dictionary – available at: https://www.lexico.com/definition/artificial_intelligence

⁸⁰ https://www.sas.com/en_gb/insights/analytics/machine-learning.html

⁸¹ See for example Chen et al. (2015) in the credit card industry

⁸² Adapted from Shobha and Rangaswamy (2018)





⁸³ It is not within the scope of this study to determine the most appropriate statistical method in this instance.

⁸⁴ For example, Google Deepmind's general purpose algorithm "AlphaZero".

Within these approaches, the selection of a particular methodology will depend not only on the context and relevant statistical criteria, but also on how the algorithm and its associated outputs are to be used. Typically, studies that implement machine learning techniques often utilise multiple learning algorithms to obtain better predictive performance (for example, random forests). These methods are known as ensemble approaches.

A1.2 Why these techniques?

With the right data and tuning, machine learning techniques may offer a number of benefits over and above traditional techniques. Some of the key potential benefits of machine learning techniques include:

-  **Accuracy and precision:** particularly within prediction and classification problems, machine learning techniques such as random forests have often been shown to outperform traditional methods such as logistic regression.
-  **Dynamic nature of machine learning:** The iterative aspect of machine learning is a key advantage; as models are replenished with updated data, they are able to independently adapt their classifications to produce repeatable decisions and results.
-  **Flexibility in processing data structures:** machine learning techniques are capable of processing large, unstructured datasets, facilitated by the increased availability and cost-effectiveness of computing power.
-  **Non-linearities and outliers:** traditional approaches are limited in their ability to process non-linearities. Through their flexibility in determining the relationship between input and output, machine learning methods are able to segment variables to more effectively account for non-linear relationships and similarly, outlying observations.

Whilst these potential benefits make the implementation of machine learning an attractive proposition, prospective users should also be conscious of the potential limitations. It may not always be the case, for example, that machine learning methods outperform traditional techniques such as logistic regression. More complex machine learning models such as deep neural networks may also be less interpretable; it may not always be immediately apparent how supervised techniques have arrived at an outcome (the “black box problem”).

As in any analytical problem, the quality of the output and the likelihood that a study is able to benefit from the implementation of machine learning methods is inherently related to the quality and granularity of the data underlying it and specific nature of the problem at hand. A more detailed discussion on these benefits and the possible trade-offs of machine learning methods can be found in Section 5.

A1.3 Example use cases

Given the potential benefits from implementing machine learning techniques, facilitated by the improved availability and cost-effectiveness of computing power, the number of machine learning use cases has grown considerably in recent years.

From film recommendation engines to fraud detection, institutions both large and small, public and private have sought to explore the use of these techniques to improve their day to day operations. Some examples of this are illustrated in Table 7:

Table 7: Example machine learning use cases

Sector	Example use cases
<i>Financial Services</i>	Banks and other financial sector firms use machine learning to develop risk assessments across their lending portfolios, fraud detection and to identify possible investment options. For example, American Express’s fraud detection system. ⁸⁵
<i>Healthcare</i>	Healthcare partners have adopted machine learning to make real-time assessments of patient health and improve analysis of patient-level data, for example. The increasing role of wearable devices has also created a number of opportunities across the general public. For example, Deepmind developed an algorithm that identifies common eye diseases from routine scans. ⁸⁶
<i>Retail</i>	Retailers have developed a large range of applications including sales forecasting, fraud detection, product recommendation engines and customer analytics. For example, the “Amazon Personalize” algorithm creates individualised recommendations for customers. ⁸⁷
<i>Public sector</i>	Government departments are increasingly seeking to utilise the benefits of machine learning in a public policy setting. In this report, examples include the use of machine learning across policing, immigration policy, homelessness prevention and long-term unemployment risk.
<i>Energy & Resources</i>	Upstream energy firms have invested in developing machine learning solutions across, for example, the discovery of new energy sources or predicting the probability of asset failures. For example, BP has invested in machine learning platforms designed to inform business decision-making and drive cost efficiencies. ⁸⁸
<i>Transportation</i>	Transport bodies have developed machine learning capability to analyse customer travel patterns and movements in key transport hubs, together with e.g. route efficiency analysis. Machine learning techniques are also utilised across demand projections and sales forecasts.

Source: Deloitte analysis

⁸⁵ <https://www.americanexpress.com/us/foreign-exchange/articles/payment-services-fraud-detection-using-AI/>

⁸⁶ <https://deepmind.com/impact>

⁸⁷ <https://aws.amazon.com/personalize/>

⁸⁸ <https://www.bp.com/en/global/corporate/news-and-insights/press-releases/bp-goes-all-in-on-aws-for-its-european-mega-data-centers.html>

Appendix 2 – Detailed case study findings

A2.1 Overview of case study findings

This section provides detailed findings for each of the four selected case studies, covering the background and objectives of the policy, methodology, key challenges, implementation, key learnings and next steps.

The following resources were used to inform the case study section, together with the stakeholder interviews. Full references can be found in the bibliography.

- **United Kingdom – Risk of reoffending:** Oswald et al. (2018); Barnes & Sherman (2019)
- **Portugal – Risk of long-term unemployment:** Bajaj et al. (2018)
- **Switzerland – Refugee resettlement:** Bansak et al. (2018)
- **USA – Risk of homelessness:** Greer et al. (2016); NYC Furman study (2017)

A2.2 United Kingdom – Risk of reoffending

Background and Objectives

Durham Constabulary's Checkpoint programme is a rehabilitation programme that provides an alternative to prosecution for a subgroup of low and moderate-level criminal offenders. The programme seeks to prevent future crime by identifying the reasons why an adult has committed an offence and by providing tailored interventions to effectively support them in desisting from crime.

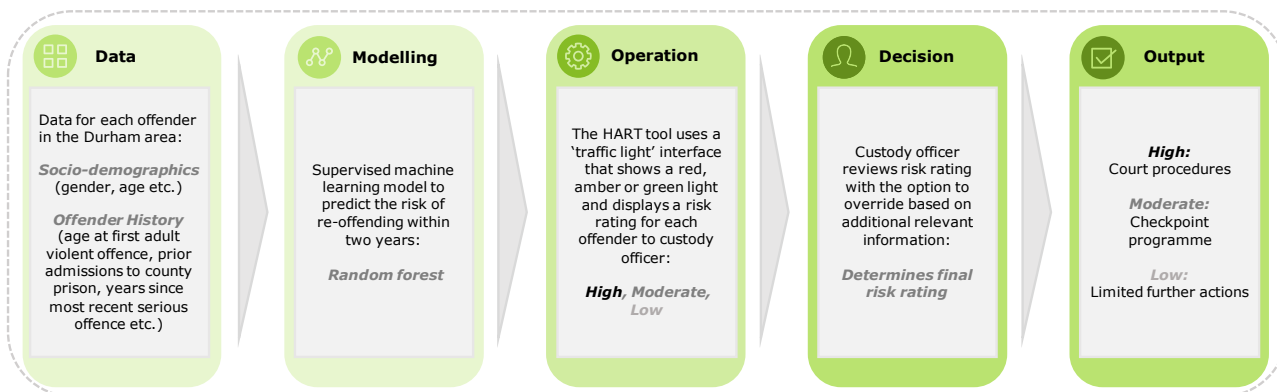
The Checkpoint programme is only available for individuals that are expected to offend within the next two years, but not in a serious violent manner. The programme not only aims to provide offenders with tailored support but also promote the efficient allocation of resources to positively impact a group of individuals who may have previously faced prosecution.

In order to identify people who are eligible for the programme, Durham Constabulary developed a decision-support technology called the Harm Assessment Risk Tool (HART), in partnership with the University of Cambridge. The HART tool was implemented in 2017 and helps to identify whether suspects are at low (unlikely to commit any crime), moderate (likely to commit a non-serious crime) or high risk (highly likely to commit a serious violent crime) of reoffending within two years. An officer subsequently acts as decision-maker, using information from other police databases combined with the recommendation from the algorithm in making the final determination.

Only people who are categorised as moderate risk are eligible for the Checkpoint Programme. The overall objective of the tool was to enable better-targeted interventions and to promote consistency in decision-making across the Constabulary.

How the HART tool operates, from data collection through to a decision, is illustrated in Figure 5.

Figure 5: HART tool process map



Source: Deloitte analysis

Methodology

Data & Matching

The machine learning model was trained by using custody data on approximately 140,000 individuals who have previously been arrested and processed in Durham over a 5 year period (2008–2012 inclusive). In total, the algorithm uses 34 different predictors, including sociodemographic characteristics (for example age, gender) and information on offender history (for example age at first violent offence, years since most serious violent offence), to create a risk score.

All data included in the model is held by Durham Constabulary, and no matching was conducted to include data from other local agencies in Durham, other police force areas, or national IT systems such as the Police National Computer or the Police National Database.

Key legislation relevant for the HART tool are the Data Protection Act and the Law Enforcement Directive,⁸⁹ an EU legislation for the processing of personal data by controllers for law enforcement.

Model / Results

Durham Constabulary uses a supervised machine learning technique (random forest) with 509 separate regression trees to predict the risk of reoffending within two years for an individual.

A value judgement is built into the model to minimise “false negatives”; the HART tool is programmed to intentionally favour cautious errors (overestimation of the risk level), to dangerous errors (underestimation of the risk level), such that the model produces approximately two cautious errors for each dangerous error. As a result, the model is particularly effective at correctly classifying those at low risk, however, this also means that a sizable proportion of high-risk forecasts are intentionally inaccurate.

An independent validation study of the HART tool was conducted in 2016 to test the overall accuracy of the algorithm with data that was not used to build the model. The validation used data on c.15,000 custody events in 2013, and compared the forecast for each of those custody events with the actual, known outcomes over the following two years.

The study found a validated accuracy of 62.8%, representing a drop from the construction accuracy of 68.5%. However, the algorithm was particularly effective at minimising instances of the most dangerous errors, whereby an offender is forecast as low risk but subsequently commits a serious

⁸⁹ Law Enforcement Directive (2016). Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/711219/LED_Document.pdf

offence (high-risk behaviour). Of the population forecasted as low risk, only 2.4% subsequently demonstrated high-risk outcomes.

Key Challenges

Practical challenges

Durham Constabulary has experienced some practical challenges regarding optimisation, the trade-off between false negatives and positives and the interface design of the tool.

Challenge	Challenge detail / mitigation
<i>Optimisation</i>	<p>A challenge for any machine learning implementation is to determine how the model should be optimised (e.g. whether overall accuracy metrics or the minimisation of particular error types are appropriate).</p> <p>Durham Constabulary, together with researchers from the University of Cambridge, decided to build a value judgement into the model, which results in a ratio of approximately two cautious errors (overpredictions) for every dangerous error (underpredictions).</p> <p>As a consequence, a large proportion of high-risk forecasts are inaccurate (in the validation study, 52.7% of those who displayed high risk behaviour were forecast as high risk in validation), however the model is particularly accurate amongst those predicted to be low risk.</p> <p>This represents an active judgement that minimising dangerous errors is of utmost importance, and consequently tailoring the optimisation criteria to meet this policy objective.</p>
<i>Predictive Power</i>	<p>The HART tool uses only data held by Durham Constabulary, which may limit the predictive power of the model.</p> <p>Early results of the validation study show a discrepancy between decisions made by the algorithm and by officers, which may be due to higher information availability (e.g. the national police databases or data on interactions with the individual that are available to the officer).</p>
<i>Tool Design</i>	<p>A key challenge relates to the design of the tool interface and how this can be best tailored to facilitate engagement with decision-makers.</p> <p>The HART tool currently shows a "traffic light" colour scale to indicate the different risk scores. Although officers cannot see which factors the algorithm used to make the decision, the risk classification is accompanied by guidance which states how the officer should consider the additional information and databases available to them in acting as decision-maker.</p>

Ethical challenges

During the implementation of the HART tool, a number of ethical considerations were made regarding transparency, discrimination and potential biases of the algorithm.

Challenge	Challenge detail / mitigation
<i>Transparency</i>	<p>Durham Constabulary has been open about the usage of the tool and the features that are incorporated into the model. A number of published research papers discuss the performance of the model together with the key learnings regarding the associated legal and ethical challenges.</p>
<i>Inconclusive evidence leading to unjustified actions</i>	<p>As set out above, the HART tool can only use data held by the Constabulary and does not include data from national police databases or other sources. As a result, the model cannot account for all potentially relevant factors that determine the risk of reoffending. This is a necessary ethical consideration as decisions based on model outcomes alone could therefore be founded on incomplete information, and may not be justifiable as a result.</p> <p>This was a core reason why the Constabulary adopted machine learning as a decision aid only, with an officer retaining decision-making responsibilities. In making a decision, the officer can access these wider information sources to inform the final classification.</p>
<i>Biases</i>	<p>A number of considerations were made regarding both data and algorithmic biases. For example, if particular subgroups have been disproportionately targeted by police action in the past, the algorithm may incorrectly assess the risk of reoffending for those individuals in future.</p> <p>Further considerations were also made regarding the use of regional indicators (postcode) within the initial model, which may be indirectly related to measures of community deprivation. Moreover, if police forces respond by focusing on the highest risk areas, this would lead to more from these areas being arrested compared to lower risk areas, potentially generating a feedback loop of increased police attention as the model is updated for these outcomes.</p> <p>As a result, the HART tool has been subject to ongoing testing and validation. Many of the parameters are flexible (e.g. the optimisation criteria), allowing the model to be updated in future should new trends emerge.</p>
<i>Intentional overprediction</i>	<p>As set out in the methodology summary, Durham Constabulary actively decided to minimise underpredictions of the risk within the optimisation criteria. This results in improved accuracy for low risk predictions, but also results in a number of inaccurate high risk predictions.</p> <p>In making these considerations, the Constabulary considered the overall benefit to society from this value judgement compared to the potential negative consequences on particular individuals.</p> <p>For example, an offender being inaccurately classified as high risk may result in them being ineligible for the Checkpoint programme, and subject to normal court procedures. However, the model also serves to protect the public from the most “dangerous” types of error through the same value judgement. From an ethical perspective, deliberately overestimating may be seen as unreasonable by some, but also a necessary side effect of protecting the public by others.</p>

Legal challenges

The implementation of the HART tool is required to comply with the Law Enforcement Directive, an EU legislation for the processing of personal data by data controllers for law enforcement purposes. The European Convention on Human Rights Article 8 has further been noted as one of the key legal frameworks.⁹⁰

Challenges	Challenge detail / mitigation
<i>European Convention on Human Rights</i>	In order to address ECHR Article 8 "Right to respect for private and family life", Durham Constabulary had to ensure a fair balance between the rights of each individual and the benefits for the wider society (referred as "experimental proportionality"). Benefits of the HART tool may refer to better outcomes for society and more consistent decision-making across custody officers.
<i>Requirements relating to automated processing</i>	The HART tool is used for advisory purposes with an officer retaining full decision-making discretion. Decisions are not based automatically on model outcomes; the HART tool is just one input into an officer's determination of risk.

Implementation

Durham Constabulary followed a transparent approach in implementing the HART tool; several published research papers have discussed the methodology, results and learnings from the tool. Independent model validation studies were undertaken in 2013 and 2019 to evaluate the model's stated accuracy. The Constabulary also ensures that the HART tool is regularly refreshed with more recent data.

The HART tool uses a 'traffic light' interface, where officers are shown a red, amber or green light for low, moderate or high risk respectively. In addition, the interface displays a reminder that the HART tool aims to assist and supports an officer's decision, but other relevant and available information such as the national police database should be used to ensure that an appropriate disposal option is given. Officers are not provided with the factors that have led to the risk classification output.

Individuals are informed if they are subsequently deemed eligible for the Checkpoint Programme based upon the officer's decision. A formal challenge mechanism relevant to model outcomes is not in place as the HART tool is only one part of the decision-making process; police forces also have a range of out of court disposal options at their discretion.

The learnings from the HART tool also informed the development of the 'ALGO-CARE framework'⁹¹ by Oswald et al. (2018) which aims to translate the ethical and legal considerations into practical steps that public bodies can take to mitigate the associated challenges. The 'ALGO-CARE' framework has been adopted by the National Police Chiefs Council Data Group and serves as a recommended guidance for police forces that consider the implementation of machine learning tools.

⁹⁰ ECHR Article 8. Available at: https://www.echr.coe.int/Documents/Guide_Art_8_ENG.pdf

⁹¹ Details of the ALGO-CARE framework are set out in section 3.2 and in Appendix 3.

Key Learnings / Next steps

Key Learnings

- It is recommended that **machine learning models can act only as a decision aid, not as a decision-maker**. Not only do models contain inaccuracies but they are also unable to account for all relevant factors that determine the target variable.
- Model outcomes should **only be one part of the decision-making, other available information should further be considered** within the final policy decision.
- It is **important to consider the context in which the algorithm is used and its desired objectives** in determining how the model should be **optimised** and subsequently **operationalised** into a policy tool.
- Durham Constabulary have adopted a **transparent approach in setting out the inputs, outputs and usage of the model, together with the key challenges and associated mitigations**. The Constabulary conducted several validation studies to evaluate the performance of the model.

Next Steps:

The HART tool is fully implemented at Durham Constabulary. Since the independent validation of the model, the Constabulary has done further work in better understanding ethical issues and support other police forces that are interested in using machine learning techniques in policing.

A2.3 Portugal – Risk of long-term unemployment

Background and objectives

Long-term unemployment (LTU), which is defined by Eurostat as a period of 12 consecutive months or longer of being unemployed, is a prevalent issue in Portugal.

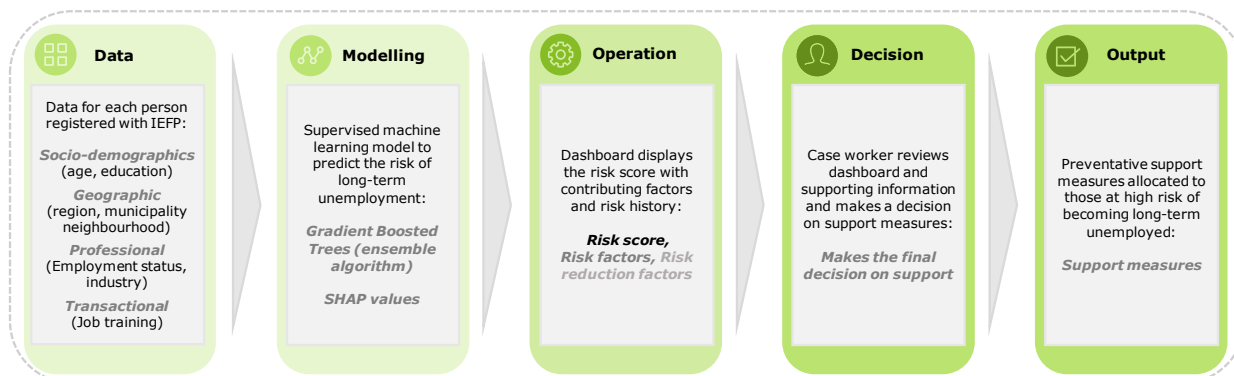
Since 2018, the Portuguese National Institute of Employment and Professional Development has developed an analytical tool that helps to predict the risk (low, medium, high) of becoming long-term unemployed for each citizen that is registered with the institution, in partnership with Nova School of Business and Economics.

IEFP can use the model outcomes to allocate resources more effectively and provide tailored services based on risk scores. The model is operationalised as a decision-support tool (DST) within a human-in-the-loop system that allows case workers to determine whether the rating is appropriate (i.e. the case worker can override the decision made by the algorithm) and the resulting resources that should be offered.

The institution previously used a logistic regression model to create the risk scores. The overall objectives of the scheme were to introduce a dynamic approach that improves the accuracy of the model and the preventative support on offer. At the time of research, the machine learning tool has been developed and is in a pilot stage. A final decision on the design of the interface and how to use model outcomes will be made in the upcoming months.

The operation of the decision-support tool is illustrated in Figure 6.

Figure 6: Long-term unemployment support tool process map



Source: Deloitte analysis

Methodology

Data & Matching

The model uses data from IEFPP, including the professional background and sociodemographic profiles of 3.5 million people registered between 2007 and 2017, along with transactional records regarding their interaction with the institution.

Data from PORDATA (a national database containing regional statistics) and the 2011 census was used to obtain further information on socioeconomic indicators at the municipality level. In addition, publically available macroeconomic indicators are incorporated into the model to represent economic events. Data is consistently updated and is anonymised such that subjects are not identifiable.

Examples of the features included in the model are:

- *Demographic*: age, gender, nationality, education
- *Geographic*: region, neighbourhood, municipality etc.
- *Professional*: Employment status, current industry, social welfare etc.
- *Transactional*: job training, job offers received, job offers declined etc.

Model / Results

The research team tested different model approaches when developing the decision-support tool. XGBoost, a decision tree-based ensemble machine learning algorithm, demonstrated the best performance in terms of precision compared to other techniques such as random forest or logistic regression. Shapley Additive Explanation (SHAP) values, a form of additive feature attribution method, were used to further develop model outputs; these explain a model’s factors for each individual as opposed to simply overall feature importance for the whole population. This has enabled the development of a dashboard that sets out the factors that contribute to an individual’s risk score.

In operationalising the outputs, trials are currently being undertaken that assess how case workers react to the output of the decision-support tool, with the objective to determine how best to facilitate engagement.

Key Challenges

Practical challenges

The development and implementation of the decision-support tool to predict the risk of long-term unemployment faces a number of practical challenges regarding optimisation, resourcing, technical validation and the creation of business rules.

Challenge	Challenge detail / mitigation
<i>Optimisation</i>	<p>A key challenge in effectively implementing the algorithm is the policy decision on what to optimise (e.g. the job that lasts the longest / pays the most / gets a person out of unemployment quickest).</p> <p>The optimisation question may depend on individual circumstances and cannot be generalised. In addition, the definition of LTU differs across countries, which makes it difficult to decide on support measures and benchmark to other countries.</p> <p>In this instance, it would be possible to adapt the definition of the target variable (e.g. the number of years considered) and optimisation criteria.</p>
<i>Resourcing</i>	<p>The potential trade-off between the increased identification of people that are at high risk of becoming long-term unemployed and the resources that IEFP can dedicate to those cases has been noted as an ongoing challenge.</p>
<i>Business rules</i>	<p>Currently, case workers are provided with the outcomes of the model, but may not understand how to translate those outcomes into a decision on support measures. A recommender system is currently under development that suggests certain actions based on different outcomes.</p>
<i>Technical validation</i>	<p>Randomised Control Trials, where the algorithm would provide LTU scores to a “treated” subsample to compare performance against current policy measures, could not be undertaken given legal considerations, in particular the responsibility to treat all subjects in the same manner.</p> <p>Standard testing and validation using techniques such as temporal cross-validation has therefore been undertaken to analyse the performance of the algorithm.</p>

Ethical challenges

A number of ethical challenges were considered within the implementation of the tool, including the potential for systematic biases reflected in the original data:

Challenge	Challenge detail / mitigation
<i>Transparency in decision-making</i>	<p>Case workers who use the outcomes of the model should be able to understand how the algorithm makes a decision; this otherwise represents a risk that decisions are made without the supporting evidence being fully understood.</p> <p>IEFP and Nova SBE are currently researching how best to inform and design model outputs to improve engagement with case workers, recognising that there may not only be limited awareness of machine learning tools, but also differences in how humans and algorithms approach risk.</p>

Challenge	Challenge detail / mitigation
<i>Biases</i>	<p>There is a risk that systematic biases may be reflected in the original data, particularly against age, gender and disability status. For example, long-term unemployment is especially prevalent amongst older workers.</p> <p>To mitigate or assess the extent of biases, protected characteristics are kept in the model to allow for bias auditing. Error assessments are able to identify which groups are discriminated against and how often this occurs.</p>
<i>Discrimination</i>	<p>The potential biases described above can lead to discriminatory outcomes against certain subpopulations. Researchers suggested that the model should achieve parity on false negatives and false positives, across protected categories (i.e. age, gender, disability status).</p> <p>Parity on false negatives would ensure that all individuals who require support measures are not systematically discriminated against. Parity on false positives would mean that no particular subgroup is disproportionately led towards more intensive preventative support measures than they require.</p>

Legal challenges

IEFP has implemented the machine learning model as a decision-support tool within a human-in-the loop system and uses an “explanation framework” to help individuals to understand how LTU risk scores are created; this serves to mitigate some of the standard legal challenges associated with machine learning, particularly those around automated processing.

Challenge	Challenge detail / mitigation
<i>Requirements for automated processing</i>	<p>The LTU risk scores created by the algorithm are only a component of the decision-making, and not a fully automated system. It is a decision-support tool that helps case workers in making decisions on support measures. Case workers can also override the suggestion of the algorithm.</p>
<i>Right to inform</i>	<p>The research team implemented an “explanation framework” to help individuals understand what factors contribute to their LTU risk score, i.e. factors that can increase and decrease their score. This in part sought to align with GDPR standards, in particular the “right to explanation”.</p>
<i>Data processing</i>	<p>Data protection regulations regarding data sharing and matching were not inhibitive in this instance; data included in the model was anonymised and already held by IEFP, and provided to Nova SBE through data sharing agreements.</p>

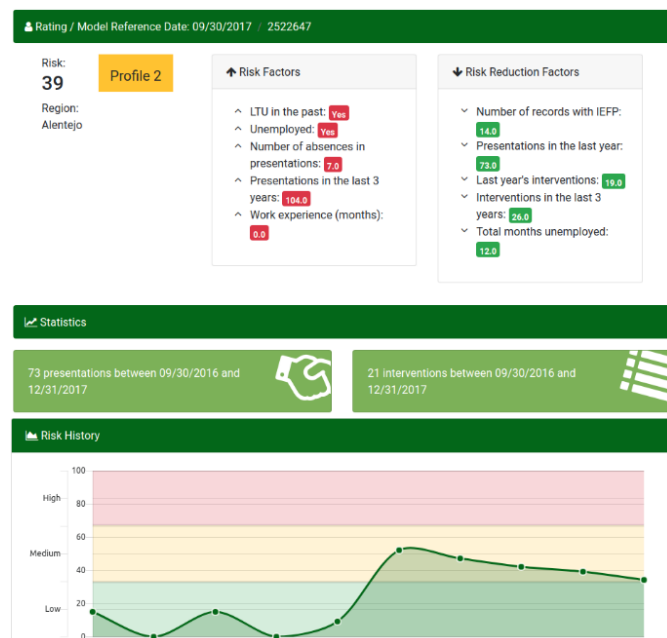
Implementation

The machine learning tool has been developed and is in the pilot stage. IEFP case workers can see a dashboard displaying the risk score for each person together with contributing factors, and their risk history (Figure 7).

Further implementation steps relate to the final design of the interface and how to foster interaction between case workers and the output of the tool. IEFP and Nova SBE are seeking to implement a recommender system, where the tool provides a recommended action to a case worker based on the outcome of the algorithm.

The machine learning tool seeks to help case workers to make an informed judgement and not to take away their authority in making a decision. A key learning from the research represents the importance of how results and recommendations are presented to decision-makers. IEFP and Nova regularly interact with case workers to develop the design of the tool and receive feedback on usage. This reflects the importance of fostering trust in the outputs of the system and to encourage their engagement with the tool.

Figure 7: Dashboard of IEFP’s decision-support tool



Source: Bajaj et al. (2018)

Key Learnings / Next steps

Key Learnings

- **Government departments should consider their policy objectives and how model outputs are to be used** throughout the development phase, **in terms of both optimisation and interface design.**
- Decision-making should be **human-centric**; model outcomes should be **used to inform, not replace, the human decision-making process.**
- **Communication and regular interaction with case workers** to develop the tool interface can facilitate trust and engagement with the outputs of the algorithm.
- The progress to date represents a step towards a more **effective decision-making and resource allocation.** It has created a **positive cultural change at IEFP regarding machine learning and artificial intelligence tools** and created the necessary infrastructure for future analysis.

Next Steps:

A number of research papers are likely to be published throughout 2020, with focus on the following:

- Business rules: how to translate model outputs into a final decision on support measures, and the initial development of a recommender system.
- Findings of the trials regarding interactions with IEPF case workers.
- Definition of the different risk classifications (low, moderate, high) and how to treat people that fall into those categories.

A2.4 Switzerland – Refugee resettlement

Background and objectives

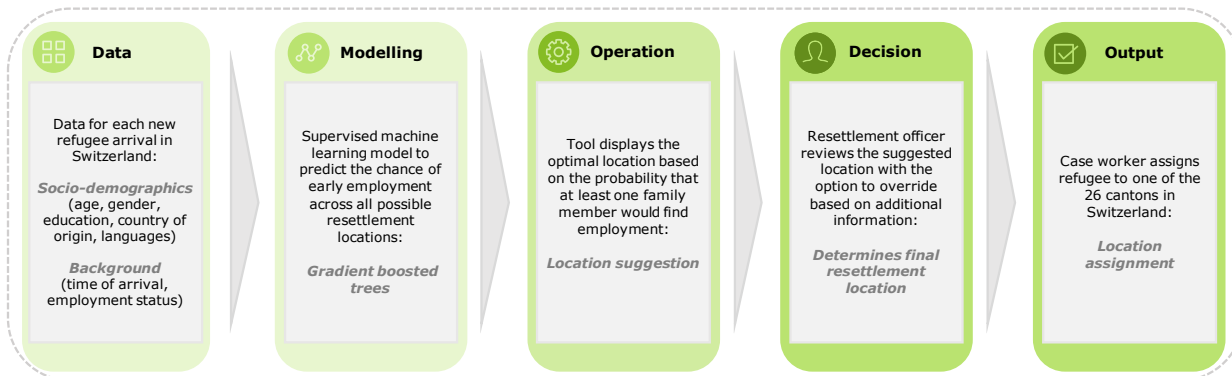
The Swiss State Secretariat for Migration developed an algorithm that seeks to improve the refugee resettlement process, in partnership with the Immigration Policy Lab at ETH Zurich and Stanford University. The overall objective is to help refugees to become integrated into society by finding employment more quickly.

In order to find the optimal resettlement location for each refugee, the algorithm predicts the probability of employment at each of the 26 Swiss cantons, and optimises at the family level to deliver a recommendation (by recommending the canton which maximises the average probability that at least one refugee in each family finds a job).

The predictive tool has been trialled in Switzerland since 2018 and is operationalised as a decision-support tool that provides placement officers with recommendations on an optimal location. The placement officer makes the final reallocation decision and can override the algorithm based on expertise or other information.

Before the introduction of the machine learning tool, the SEM assigned refugees randomly to the 26 cantons based on proportional distribution. Switzerland is one of the first countries that has trialled a data-driven tool in the refugee context. The operation of the tool is illustrated below in Figure 8.

Figure 8: Refugee allocation process map



Source: Deloitte analysis

Methodology

Data & Matching

The algorithm draws on data of 22,159 refugees from the ZEMIS database, covering the period 1999 to 2013. The ZEMIS database is held by the SEM to process asylum claims and record employment information for refugees. The predictive model incorporates data on refugees’ socioeconomic characteristics, geographical context and the synergies between them. Examples include:

- *Socioeconomic characteristics:* age, gender, country of origin, language skills, education.
- *Background information:* time of arrival, assigned location, measured employment status.

Model / Results

A number of supervised machine learning methods were used to predict the chance of employment for refugees across all possible resettlement locations. Gradient boosted trees have shown the greatest performance over other methods such as random forest, elastic-net logistic regression, and kernel-based regularized least squares.

A separate model was used for subgroups of refugees assigned to each location, allowing for the discovery of synergies between location and refugee characteristics. Refugee-level predictions were transformed to the case-level as refugees are usually assigned on a family rather than an individual basis. The algorithm is flexible and can accommodate several criteria and constraints such as an assignment restriction (i.e. the maximum number of refugees that can be sent to each canton).

Out of sample test results have shown that the third year employment rate for refugees was 26% when using the algorithm for resettlement decision, compared to 15% under the previous system of random distribution. Evidence suggests that the predictive tool has the potential to increase the third year employment rate by approximately 73% in Switzerland.

Key challenges

Practical challenges

At the current stage, this example has encountered challenges that mainly relate to the evaluation of the scheme.

Challenge	Challenge detail / mitigation
<i>Validation</i>	Evaluation results have shown that the predictive tool has the potential to increase refugees' third-year employment rate by about 73% in Switzerland. However, the "real world" impact of the tool can only be measured after a 2-3 year-long pilot programme.

Ethical challenges

Ethical considerations may relate to potential fairness concerns, for example, that benefits are not achieved across all refugees and locations.

Challenge	Challenge detail / mitigation
<i>Transparency</i>	In order to address concerns regarding transparency, the research team has published the source code and a supplementary document that describes in detail the methodology of the model and the data that is used.
<i>Fairness</i>	<p>The model is unable to capture refugee preferences. However, preference-based matching is effective only under the assumption that these preferences are well-informed.</p> <p>A non-preference based approach avoids the possibility that communities develop preferences regarding certain refugee characteristics. The algorithm is also flexible in programming constraints into the allocation mechanism that align with distribution quotas across the 26 cantons.</p>

Legal challenges

The use of personal data has to comply with the Swiss Federal Data Protection Act, a revised version of which is expected to be passed in 2020.⁹² Its provisions are similar to those of GDPR, although with some conceptual differences.

Challenge	Challenge detail / mitigation
<i>Requirements for automated processing</i>	The predictive tool has been implemented as a decision-support tool, and not as a decision-maker, and therefore does not rely solely upon automated processing. The tool provides placement officers with the recommended location for each case. However, the placement officer can decide whether to accept the recommendation provided by the algorithm or to override it.
<i>Data sharing and matching</i>	The algorithm only uses data already held by SEM, and as such has not encountered challenges with sharing and matching across multiple sources.

Implementation

The State Secretariat for Migration first implemented the machine learning tool for a 6 month period across 2018-2019 as a technical pilot. At the end of the pilot, the Swiss government announced the introduction of a new allocation system and the model was adjusted accordingly.

The SEM introduced a second pilot of the updated model in January 2020. The pilot is based on a Randomised Control Trial, where 1000 refugees are allocated to the 26 cantons based on their highest probability of early employment. The pilot will run for approximately 2-3 years. At the end of the pilot, it will be possible to conduct an empirical evaluation of the impact and cost-effectiveness of the scheme over and above the existing random assignment mechanism.

Key Learnings / Next steps

Key Learnings

- Learnings from the refugee assignment case suggest the use of a **human-centric decision-making process, informed by machine learning tools.**
- An **effective implementation strategy should include ongoing testing and validation prior to formal implementation**
- **A transparent approach can seek to mitigate reputational risk** (e.g. by making the code and research papers publically available).
- The predictive tool has the potential to increase refugees' employment rate in Switzerland by **improving the previous system of random assignment.**

Next Steps:

The next steps of the scheme are highly dependent on the pilot evaluation results. However, as the Swiss government has been open to machine learning techniques and the pilot of the scheme, it is expected that Switzerland will continue using the tool in the future if this is shown to be successful.

⁹² <https://www.edoeb.admin.ch/edoeb/en/home/documentation/annual-reports/older-reports/12th-annual-report-2004-2005/revision-of-the-federal-data-protection-act.html>

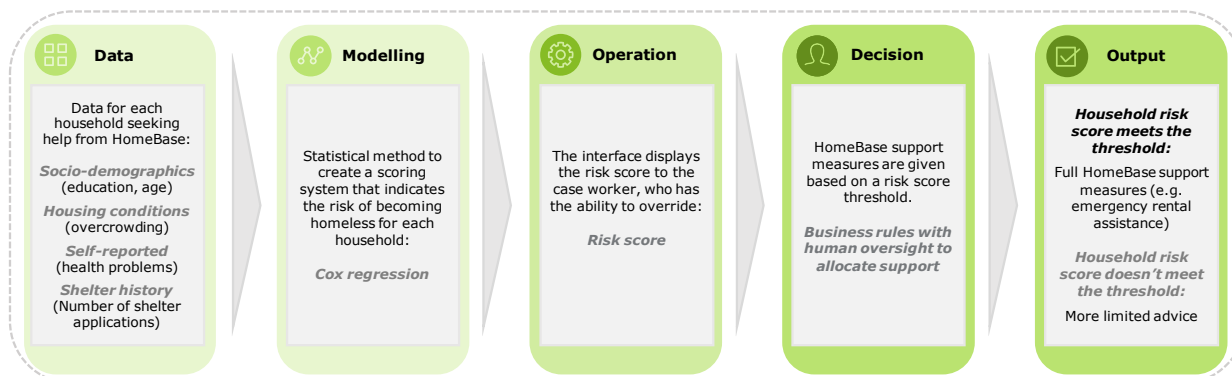
A2.5 United States – Risk of becoming homeless

Background / Objectives

The Department of Homeless Services introduced the HomeBase programme in 2004 to address the issue of homelessness in New York City; responsibility for the scheme now falls within the remit of the Human Resources Administration. The City works with 7 community non-profit organisations to provide preventive services to households at risk of becoming homeless. The non-profit organisations are split geographically to administer the HomeBase scheme and follow a standardised way to provide the service. About 28,000 families are served by the programme each year.

The allocation of support involves a scoring system that measures the risk of becoming homeless. Households receive HomeBase services based on their risk score, including emergency rental assistance, access to job training and landlord mediation services. At this stage, the method employed is purely statistical, however the outputs of the scoring system directly translate into the support offered to families. The decision-making process is illustrated in Figure 9.

Figure 9: HomeBase decision support - process map



Source: Deloitte analysis

Furthermore, HRA plans to implement predictive tools in the future to improve the outreach of the HomeBase programme, with particular focus on identifying those that are at the highest risk of becoming homeless. This could seek to adopt research conducted by the Centre for Innovation through Data Intelligence (CIDI), in partnership with the New York University's Furman Centre, which used data on human services, buildings and neighbourhoods to predict families' risk of homelessness by using a supervised machine learning technique.

Methodology

Data & Matching

Information collected to develop the scoring mechanism included sociodemographic characteristics, housing conditions and shelter history. In order to complete missing data for factors such as young children or pregnancy, c.80% of applicants were matched with welfare case records.

The following variables were included in the Cox Regression model that informed the scoring system:

- *Sociodemographic characteristics:* gender, English speaker, number of children, education.
- *Housing conditions:* overcrowding, eviction threat, rent, currently receiving a subsidy.
- *Self-reported circumstances:* chronic health problems, criminal justice involvement, domestic violence, adolescent mother.

- *Shelter history*: shelter history as an adult, number of previous shelter applications, etc.

Model / Results

The scoring system enables HRA to provide preventive services to those that are at the highest empirical risk of becoming homeless and also enables case workers the opportunity to partially override the results of the system based on other considerations. Those seeking support are interviewed on presentation, with the data collected used to inform the decision regarding the support offered.

Evidence suggests that selecting households based on model outcomes rather than case workers' judgement alone has improved the provision of homeless services: one study (Greer et al. 2016) that tracked over 10,000 individuals subsequently improved shelter entry predictions by 77% and reduced unidentified cases of subsequent homelessness by 85%.

Key challenges

Practical challenges

Scheme awareness has been noted as one of the key practical challenges of the HomeBase programme. Further challenges refer to resource availability to implement new technologies within the scheme and technical validation.

Challenge	Challenge detail / mitigation
<i>Awareness</i>	A key challenge has been raising awareness of the scheme and engaging with the households most in need. Recent policies include the deployment of mobile vans and targeted letters to increase the awareness of the scheme across areas that may have a large number of high-risk families. These interventions have been successful in reaching out to families that would not have registered with the scheme.
<i>Resourcing</i>	HRA is planning to introduce machine learning methods to improve the outreach of the HomeBase programme. A key challenge has been the availability of resources to develop and implement these new solutions.
<i>Reputational risk</i>	A Randomised Control Trial has previously been commissioned (i.e. measuring the impact of the scheme by analysing areas where the scheme is in place and areas without the scheme). This study led to criticism by politicians regarding the fairness of the scheme (i.e. people in some areas are excluded from support). However, at the time DHS noted that every person had the possibility to go to one of the partner agencies and receive support regardless of their location.

Ethical challenges

Ethical challenges in this instance particularly relate to transparency and fairness:

Challenge	Challenge detail / mitigation
<i>Transparency</i>	Several research papers have been published including the methodology of the scoring system and how outcomes are used in practice, which contribute to increased transparency.
<i>Discrimination</i>	HRA has faced only limited ethical challenges with the implementation of the statistical method as the scoring system is being used for preventative support (as opposed to classification of an individual in a particular state

such as being “homeless”). However, HRA has received some complaints from households that were not eligible for certain HomeBase services based on their risk score which the agency addressed by making service referrals that matched the need or re-evaluating the circumstances of the case.

Legal challenges

The United States does not have to comply with GDPR legislation. However, New York City has to comply with state-level data protection regulation such as the Electronic Data Security Act (SHIELD)⁹³ and did encounter some challenges regarding data sharing.

Challenge	Challenge detail / mitigation
-----------	-------------------------------

<i>Data sharing</i>	Data protection regulations regarding privacy represented a barrier to obtain a wider range of datasets such as information from the child protection service.
---------------------	----------------------------------------------------------------------------------------------------------------------------------------------------------------

Implementation

HRA is currently using the scoring system and has business rules in place that determine the level of support based on the risk scores. At a certain threshold, families are provided with the full HomeBase support, those below receive more limited advice. Families with children have different eligibility criteria and thresholds than families without children, for example. Moreover, households who receive only limited support have the possibility to receive more support in the future should their circumstances change.

Although the system provides clear business rules, case workers have the ability to override model decisions in some circumstances to reflect unique household circumstances.

Key Learnings / Next steps

Key Learnings

- Tools and techniques **face reputational risk regardless of their methodology, and regardless of whether schemes are seeking only to allocate benefits.**
- Statistical tools can be used **both for classification problems and in prediction for the purposes of allocating preventative support.**
- The **resources required to implement machine learning tools can be burdensome** for organisations that are already fully utilised and may not have in house expertise.
- To realise the benefits from quantitative techniques in a policy setting, **the implementation of statistical methods should be complemented by practical measures to raise awareness and scheme uptake.**

Next Steps:

HRA is seeking to explore the use of machine learning techniques within the HomeBase programme, similar to the CIDI and Furman Institute study. The overall objective is to introduce a dynamic model that improves outreach and the allocation of preventative support.

⁹³ Available at: <https://www.nysenate.gov/legislation/bills/2019/s5575>

Appendix 3 – The ALGO-CARE framework

A3.1 The ALGO-CARE framework – learnings from UK policing

As introduced in section 3.2, the ALGO-CARE framework developed by Oswald et al. (2018) reflects the experiences of the Durham Constabulary in developing the HART tool. The framework brings together key considerations in terms of practical, legal and ethical challenges into a machine learning decision-making framework that is transferable to other use cases across the public sector.

Table 8 below sets out the ALGO-CARE mnemonic, as set out in Oswald et al. (2018) pp245-248.

Table 8: The ALGO-CARE mnemonic

Consideration	Key questions
<i>Advisory</i>	<p>Is the assessment made by the algorithm used in an advisory capacity?</p> <p>Does a human officer retain decision-making discretion?</p> <p>What other decision-making by human officers will add objectivity to the decisions (partly) based on the algorithm?</p>
<i>Lawful</i>	<p>On a case-by-case basis, what is the policing purpose justifying the use of the algorithm, both its means and ends?</p> <p>Is the potential interference with the privacy of individuals necessary and proportionate for legitimate policing purposes?</p> <p>In what way will the tool improve the current system and is this demonstrable?</p> <p>Are the data processed by the algorithm carefully lawfully obtained, processed and retained, according to a genuine necessity with a rational connection to a policing aim? Is the operation of the tool compliant with national guidance?</p>
<i>Granularity</i>	<p>Does the algorithm make suggestions at a sufficient level of detail / granularity, given the purpose of the algorithm and the nature of the data processed?</p> <p>Is data categorised to avoid 'broad-brush' grouping and results, and therefore issues potential bias?</p> <p>Do the benefits outweigh any technological or data quality uncertainties or gaps?</p> <p>Is the provenance and quality of the data sufficiently sound?</p> <p>Consider how often the data should be refreshed. If the tool takes a precautionary approach towards false negatives, consider the justifications for this.</p>

Consideration Key questions

<p><i>Ownership</i></p>	<p>Who owns the algorithm and the data analysed?</p> <p>Does the force need rights to access, use and amend the source code and data analysed?</p> <p>How will the tool be maintained and updated?</p> <p>Are there any contractual or other restrictions which might limit accountability or evaluation?</p> <p>How is the operation of the algorithm kept secure?</p>
<p><i>Challengeable</i></p>	<p>What are the post-implementation oversight and audit mechanisms e.g. to identify any bias?</p> <p>Where an algorithmic tool informs criminal justice disposals, how are individuals notified of its use (As appropriate in the context of the tool's operation and purpose)?</p>
<p><i>Accuracy</i></p>	<p>Does the specification match the policing aim and decision policy?</p> <p>Can the stated accuracy of the algorithm be validated reasonably periodically?</p> <p>Can the percentage of false positives/negatives be justified?</p> <p>How was the method chosen as opposed to other available methods? What are the consequence of inaccurate forecasts? Does this represent an acceptable risk (in terms of both likelihood and impact)?</p> <p>Is the algorithmic tool deployed by those with appropriate expertise?</p>
<p><i>Responsible</i></p>	<p>Would the operation of the algorithm be considered fair?</p> <p>Is the use of the algorithm transparent (taking account of the context of its use), accountable and placed under review alongside other IT developments in policing?</p> <p>Would it be considered to be for the public interest and ethical?</p>
<p><i>Explainable</i></p>	<p>Is appropriate information available about the decision-making rule(s) and the impact that each factor has on the final score or outcome (in a similar way to a gravity matrix)?</p> <p>Is the force able to access and deploy a data science expert to explain and justify the algorithmic tool (in a similar way to an expert forensic pathologist)?</p>

Source: Direct quotation of Oswald et al. (2018) pp245-248

Appendix 4 – Estimation of benefits

A4.1 Benefit estimation – worked example

This section presents an assumption-based example to illustrate how improved identification due to machine learning could benefit both fuel poor households and also society at large.

This hypothetical example sets out an approach to illustrate the benefits that could result from improved identification. It is not indicative of the scale of benefits that could be expected in practice; further work is required to develop a formal cost-benefit analysis that considers the implementation of machine learning techniques. Moreover, how the provision of targeted support might be adapted in future is a matter for scheme administrators; this analysis cannot speculate as to how this would be achieved in practice. The figures set out in this section should be considered together with these limitations.

ECO3 is selected as a reference point within the example; not only was this policy developed with a greater focus on the fuel poor than in previous iterations of the scheme, but there also exists greater scope for the use of machine learning tools in this instance relative to other initiatives that are already fully or mostly automated (e.g. Winter Fuel Payment). This example, therefore, considers a hypothetical future iteration of the scheme that is able to realise the benefits from using machine learning, taking into account existing budget constraints.

Machine learning can facilitate an understanding of both of the distribution of fuel poor households and also of the factors that influence whether a household is in practice fuel poor. These insights can be used not only to improve the identification of fuel poor households to benefit existing schemes and reduce search costs, but also to tailor the design of support schemes in future. In the case of ECO, the fuel poor households accepting support can subsequently benefit from improved energy efficiency and resulting bill savings.

Under an assumption that this mechanism results in 25% of fuel poor households that are currently ineligible for ECO3 becoming eligible within a hypothetical future iteration of the scheme (and a corresponding number of non-fuel poor, eligible households becoming ineligible to reflect fixed budget constraints) this hypothetical example would result in:

- An additional c.335,000 fuel poor households being identified for support, of which c.100,000 accept the offer of energy efficiency installations.
- Potentially large scope for reductions in search cost; under existing policies, it is estimated that identifying the c.335,000 additional fuel poor households referenced above would cost c.£100m, indicating scope for cost efficiencies.
- Total annual bill savings of c.£20m across the c.100,000 households that accept support (£200 per household), with a net annual impact (societal benefit) of £9.7m (£97 per household) taking into account the reduced eligibility of non-fuel poor households and the relative income levels of the fuel poor relative to the median.
- A net present value of c.£143m if net annual impacts are considered over a 20-year horizon.

Assumptions and Limitations

The full list of quantitative assumptions used in this example is set out in Table 9.

Table 9: Assumptions employed

Assumption	Value	Source
Improvement in eligibility of fuel poor households	25%	Hypothetical scenario
Total households receiving support	Fixed	Fixed at current ECO3 levels from Detailed Fuel Poverty Tables, to represent fixed budget constraints
Uptake of ECO support offered	30%	Informed by the ECO3 Final Impact Assessment
Assumed annual bill savings per fuel poor household	£200	Informed by multiple sources including Which? and the Energy Saving Trust
Income proxy for fuel poor group	£13,672	BEIS Annual Fuel Poverty Statistics
Marginal utility of income	1.3	Treasury Green Book
NPV Discount rate	3.5%	Treasury Green Book
NPV measurement period (years)	20	Conservative assumption informed by Ofgem ECO3 delivery guidance
Supplier search cost per "lead"	£300	ECO3 Final Impact Assessment (analysis of mid scenario)

The following qualitative assumptions and limitations should also be considered in line with the figures presented in this case study:

- This example only considers a hypothetical increase in the eligible, fuel poor group resulting from improved identification through machine learning. It does not consider other movements, for example those driven by false negative or positive results.
- Given fixed budget constraints, it is assumed that scheme administrators have the ability to alter eligibility criteria such that currently eligible, non-fuel poor households become ineligible in this future iteration.
- It is assumed that energy suppliers offer support to the full range of households identified, of which a proportion agree to the installation of energy efficiency measure at the property.
- The example considers only the benefits from machine learning, both in terms of bill savings to fuel poor households and reductions in search costs. It does not consider any administrative or labour costs involved with the operation of the scheme.
- Calculations of net present value do not account for asset depreciation (e.g. reductions in boiler efficiency over time), which could serve to reduce the scope for benefits.

Net benefits resulting from improved identification

The following tables illustrate how the implementation of machine learning techniques can drive both bill savings for fuel poor households and net societal benefits:

1. Current eligibility for ECO3 ("Help to Heat") is taken as a starting point for improvements to existing policies, using data from BEIS's detailed fuel poverty tables.

Table 10: Fuel poor households, by eligibility for ECO3 Help to Heat Group

Number of households (000's)				
Eligible for ECO3?	Not fuel poor	Fuel poor	Total	Share of fuel poor eligible
Yes	3,472	1,194	4,666	47.2%
No	17,192	1,338	18,530	52.8%
All households	20,664	2,532	23,196	100.0%

Source: Table 34, Detailed fuel poverty tables

2. It is assumed in this hypothetical example that improved identification and scheme design, resulting from the implementation of machine learning, results in 25% of currently ineligible fuel poor households being identified as eligible for support (approximately 335,000 households).

The total number of households receiving support is held constant at current ECO3 levels to reflect fixed budget constraints. As such, it is assumed that eligibility criteria can be amended such that a corresponding number of non-fuel poor households are no longer eligible for support. This would result in the scheme composition set out in Table 11.

By shifting the additional c.335,000 fuel poor households into the eligible pool, and reallocating an equivalent number from the eligible, non-fuel poor group, the share of fuel poor households eligible for the scheme would increase from 47.2% to 60.4%

Table 11: Fuel poor households by eligibility, hypothetical future iteration – Illustrative outputs

Number of households (000's)				
Eligible for ECO3?	Not fuel poor	Fuel poor	Total	Share of fuel poor eligible
Yes	3,138	1,529	4,666	60.4%
No	17,527	1,004	18,530	39.6%
All households	20,664	2,532	23,196	100.0%

Source: Deloitte analysis

3. Of the additional fuel poor households identified as eligible for support, 30% (approximately 100,350 households) are assumed to accept the offer of energy efficiency installations at their property. This assumption is based on current ECO3 'findability' rates (i.e. the proportion of technical potential that can be identified and installed within any given year), which range from c.11% for solid wall insulation up to 100% for central heating measures.⁹⁴ A number of initiatives at the local level are currently in place to drive uptake, including not only the LA Flex mechanism but also signposting and other holistic support. A number of "managing agents" have also developed successful partnerships with Local Authorities to target households in need directly. It

⁹⁴ ECO3 Final Stage Impact Assessment. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/671111/ECO_3_Final_Stage_IA_Final.pdf

is assumed that within this hypothetical scheme, energy suppliers offer support to the full set of eligible households identified.

- Based upon publically available information regarding the potential savings resulting from measures available under ECO3 (e.g. loft insulation), it is assumed that households receiving support benefit from bill savings of £200 per year, leading to total savings across the additional fuel poor households of c.£20m in the year of first receiving support. In developing this assumption, it is noted that fuel poor households are likely to have larger energy needs relative to the general population (often given that these households are typically larger in terms of inhabitants, but also that these are in many cases families with dependent children).

Table 12: Estimation of total bill savings net of uptake assumption – Illustrative outputs

Variable	Label	Value
Additional fuel poor households identified	A	334,500
Assumed uptake rate	B	30%
Household accepting support	C = A x B	100,350
Assumed annual bill savings per fuel poor household	D	£200
Total annual bill savings	E = C x D	£20,070,000

Source: Deloitte analysis

- To approximate the net societal impact, assumptions from the Treasury Green Book are employed to assess the value of these savings for those in fuel poverty (whose incomes are proxied by the LIHC income threshold of £13,672) relative to those non-fuel poor that would become ineligible (proxied by the median income after housing and fuel costs, £22,787).⁹⁵

As set out in the Green Book: 'The basis for distributional weights is the economic principle of the diminishing marginal utility of income. It states that the value of an additional pound of income is higher for a low income recipient and lower for a high-income recipient. Broadly a value of 1 for the marginal utility of income would indicate that the utility of an additional pound is inversely proportional to the income of the recipient. An additional £1 of consumption received by someone earning £20,000 per year would be worth twice as much than to a person earning £40,000.'

A marginal utility of income assumption of 1.3 is used in line with the Treasury Green Book: this figure is used by DWP within distributional analysis and the calculation of welfare weights.⁹⁶

The relative income assumptions used in this example are likely conservative; fuel poor households may have income levels considerably below the LIHC threshold, whilst the use of unequivalised figures does not account for the typically larger size (in terms of inhabitants) of fuel poor households relative to the average, for example.

- Based on these assumptions, savings of £200 per year for a fuel poor household would be 'worth' £103 per year for a non-fuel poor household, representing an annual societal benefit of £97 per fuel poor household now in receipt of support (total annual net societal savings of c.£9.7m).

⁹⁵ BEIS Fuel Poverty Statistics report (2019), 2017 data, available at:

https://assets.publishing.service.gov.uk/Annual_Fuel_Poverty_Statistics_Report_2019_2017_data_.pdf

⁹⁶ Treasury Green Book. Available at:

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/685903/The_Green_Book.pdf

Table 13: Annual benefits to fuel poor and wider society – Illustrative outputs

Variable	Label	Value
LIHC income threshold	F	£13,672
Median incomes (after housing and fuel cost)	G	£22,787
Marginal utility of income	H	1.30
Redistributive effect	$I = (G/F)^H$	1.94
Estimation of net societal impact		
Annual bill savings per fuel poor household	D	£200
Value of annual bill saving for non-fuel poor	$J = D / I$	£102.95
Net annual benefit per fuel poor household	$K = D - J$	£97.05
Additional fuel poor households eligible	C	100,350
Net annual benefits from improved identification	$L = K \times C$	£9,739,000

Source: Deloitte analysis

7. Using Green Book assumptions on discount rate (3.5%) over a 20 year period, these annual societal savings of c.£9.7m would have a net present value of c.£143m. 20 years is selected as a conservative assumption given the likely asset lives of these measures.⁹⁷

Table 14: Net present value of accumulated savings – Illustrative outputs

Variable	Label	Value
Discount rate	M	3.5%
Measurement period (years)	N	20
Net present value of benefits to fuel poor	O	£143.3m

Source: Deloitte analysis

8. Together with potential bill savings, it is important to consider the benefits from reduced search costs that could result from machine learning; the ECO3 final stage impact assessment provides a number sensitivities in this regard.⁹⁸ This informs a search cost assumption of £300 per property under current policy measures.

On this basis, identifying the c.334,500 households found in this example would cost suppliers c.£100m in search costs under existing policy operations. Whilst machine learning would not be cost free in practice, there is clearly potential scope for efficiency savings in this regard.

Table 15: Search costs under current policy – Illustrative outputs

Variable	Label	Value
Search costs per property	P	£300
Total households identified	A	334,500
Total estimated search costs under current policy	$Q = A \times P$	£100.4m

Source: Deloitte analysis

⁹⁷ Ofgem data suggests measures such as wall insulation should last at least 25 years, if accompanied by an equivalent guarantee. https://www.ofgem.gov.uk/system/files/docs/2018/11/eco3_guidance_delivery_final.pdf

⁹⁸ ECO3 Final Impact Assessment. Available at: https://assets.publishing.service.gov.uk/government/ECO_3_Final_Stage_IA_Final.pdf

Glossary

AI	Artificial Intelligence
BEIS	Department for Business, Energy and Industrial Strategy
CDEI	Centre for Data Ethics and Innovation
CFP	Committee on Fuel Poverty (partner organisation sponsored by BEIS)
CIDI	Centre for Innovation through Data Intelligence
CWP	Cold Weather Payment
DCMS	Department for Digital, Culture, Media & Sport
DHS	Department of Homeless Services
DPA	Data Protection Act 2018
DPIA	Data Protection Impact Assessment
DST	Decision-support tool
DWP	Department for Work and Pensions
ECHR	European Convention on Human Rights
ECO	Energy Company Obligation
EHS	English Housing Survey
EPC	Energy Performance Certificate
FADP	Federal Act on Data Protection
GDPR	General Data Protection Regulation
HART	Harm Assessment Risk Tool
HITL	Human-in-the loop
HMRC	Her Majesty's Revenue and Customs
HRA	Human Resources Administration
ICO	Information Commissioner's Office
IEFP	The Portuguese Institute for Employment and Vocational Training
IPL	Immigration Policy Lab
LIHC	Low Income High Cost measure of fuel poverty
LILEE	Low Income Low Energy Efficiency measure of fuel poverty

LTU	Long-term unemployed
ML	Machine Learning
NEED	National Energy Efficiency Data
NPCC	National Police Chiefs' Council
OS	Ordnance Survey
RCT	Randomised Control Trial
SBE	Nova School of Business and Economics
SEM	State Secretariat for Migration
SHAP	Shapley Additive Explanation
SHIELD	Stop Hacks and Improve Electronic Data Security Act
VOA	Valuation Office Agency
WFP	Winter Fuel Payment
WHD	Warm Home Discount
ZEMIS	Central Migration Information System

Bibliography

The following academic papers were reviewed within this research. Links to sources from online resources can be found within the footnotes of this document:

Babuta, A. (March 2018). Innocent Until Predicted Guilty? Artificial Intelligence and Police Decision-Making. RUSI Newsbrief, Vol.38, No.2.

Babuta, A., & Oswald, M. (2020). Data Analytics and Algorithms in Policing in England and Wales: Towards a New Policy Framework. RUSI Occasional Papers.

Bansak, K., Ferwerda, J., Hainmueller, J., Dillon, A., Hangartner, D., Lawrence, D. and Weinstein, J. (2018). Improving refugee integration through data-driven algorithmic assignment. *Science* 359, pp325-329.

Barnes, G. C., & Sherman, L. (2018). Needles & Haystacks. University of Cambridge Research Horizons, Issue 35, pp32-33.

Bright, J., Ganesh, B., Seidelin, C. and Vogl, T. (2019). Data Science for Local Government. Oxford Internet Institute, University of Oxford.

Butaru, F., Chen, Q., Clark, B., Das, S., Lo., A & Siddique, A. (2015). Risk and Risk Management in the Credit Card Industry. NBER Working Paper No. 21305.

Centre for Innovation through Data Intelligence and NYU Furman Center. (2017). Research Brief - Predicting Homelessness for Better Prevention.

Goodman, S., Messeri, P., and O'Flaherty, B. (2016). Homelessness prevention in New York City: On average, it works. *Journal of Housing Economics*. Vol. 31, pp14-34.

Greer, A. L., Shinn, M., Kwon, J., & Zuiderveen, S. (2016). Targeting Services to Individuals Most Likely to Enter Shelter: Evaluation of the Efficiency of Homelessness Prevention. *Social Service Review*, Vol. 90, No.1.

Martínez de Rituerto de Troya, Í., Chen, R., Moraes, L. O., Bajaj, P., Kupersmith, J., Ghani, R., Bras, N. and Zejnilovic, L. (2018). Predicting, explaining, and understanding risk of long-term unemployment. 32nd Conference on Neural Information Processing Systems (NIPS). AI for Social Good. Montréal, Canada.

Oswald, M., and Babuta, A. (2018). Machine Learning Algorithms and Police Decision-Making: Legal, Ethical and Regulatory Challenges. Whitehall Report 3-18.

Oswald, M., and Babuta, A. (2019). Machine Learning Predictive Algorithms and the Policing of Future Crimes: Governance and Oversight. In *Policing and Artificial Intelligence* (Dr John L.M. McDaniel and Prof Ken Pease OBE eds., Routledge, Forthcoming 2020).

Oswald, M., Grace, J., Urwin, S., & Barnes, G. C. (2018). Algorithmic risk assessment policing models: lessons from the Durham HART model and 'Experimental' proportionality. *Information & Communications Technology Law*, Vol. 27, No. 2, pp223-250.

Rolston, H., Geyer, J., Locke, G., Metraux, S., & Treglia, D. (2013). Evaluation of the HomeBase Community Prevention Program. Abt Associates - prepared for the NYC Department of Homeless Services.

- Russell, S. J. & Norvig, P. (2016). *Artificial Intelligence: A Modern Approach*. Pearson Education Limited, 4th Edition.
- Shinn, M., Greer, A. L., Bainbridge, J., Kwon, J., & Zuiderveen, S. (2013). Efficient Targeting of Homelessness Prevention Services for Families. *American Journal of Public Health, Suppl. 2*, pp324-30.
- Shobha, G. & Rangaswamy, S. (2018). Computational Analysis and Understanding of Natural Languages: Principles, Methods and Applications. *Handbook of Statistics, Vol. 38*.
- Smith, N. (2005). *Understanding Family Homelessness in New York City: An in-depth study of families' experience before and after shelter*. New York City: Vera Institute of Justice.
- Van der Schaar, M., & Davies, S. (2018). Machine learning for individualised medicine. *Annual Report of the Chief Medical Officer, Health 2040 - Better Health Within Reach, Chapter 10*, Department of Health and Social Care.



Deloitte LLP is a limited liability partnership registered in England and Wales with registered number OC303675 and its registered office at 1 New Street Square, London, EC4A 3HQ, United Kingdom.

Deloitte LLP is the United Kingdom affiliate of Deloitte NSE LLP, a member firm of Deloitte Touche Tohmatsu Limited, a UK private company limited by guarantee ("DTTL"). DTTL and each of its member firms are legally separate and independent entities. DTTL and Deloitte NSE LLP do not provide services to clients. Please see www.deloitte.com/about to learn more about our global network of member firms.

© 2020 Deloitte LLP. All rights reserved.