Department for
Business, Energy
& Industrial Strategy

ANNEX A

Machine learning and fuel poverty targeting

July 2017

Contents

Introduction

The National Energy Efficiency Data-framework (NEED) is a rich dataset containing a variety of information about energy consumption in households and the various determinants that underlie this. This paper investigates the possibility of using NEED, in conjunction with modern analysis techniques, to provide new insights pursuant to the delivery of departmental objectives. As a particular test case, the potential of using NEED, in combination with machine learning, to help identify fuel poor homes in England, is investigated.

The fuel poverty targeting challenge is outlined in more detail below. A brief introduction to machine learning is then presented, followed by further details of the data used in this exercise. The modelling approach taken is then described, along with details of how the final prediction algorithm was selected and tuned. The results of the exercise are then discussed, detailing how this analysis does indeed show some potential in addressing the targeting issue. The geospatial aspects of the results are summarised in a couple of maps showing both the predicted levels of fuel poverty in different areas and also the rate (normalised by the population). Finally, the results are compared and contrasted to similar work done in this area.

The policy issue: Fuel poverty targeting

Fuel poverty (broadly, being unable to afford to keep one's home adequately heated) is a condition that affects around one in ten households in England. The Department for Business, Energy and Industrial Strategy (BEIS) has a number of policies designed to both provide support to these vulnerable households, and to help reduce the number of households in fuel poverty. The efficient delivery of these policies is dependent on establishing an effective way of identifying the relevant households, both to reduce the cost of finding them, and to minimise the possibility of support being given to the wrong households.

The NEED dataset contains a number of indicators that might be informative on the fuel poverty status of households, which could potentially act as proxies in a model that predicts the fuel poverty status of households (in the absence of full data that would allow a definitive categorisation). Machine learning (introduced below) has emerged as a leading contender in the field of prediction, making it an ideal candidate to explore as a helpful approach to the fuel poverty targeting challenge.

What is machine learning?

Machine learning is a subfield of computer science that gives computers the ability to learn without being explicitly programmed (Arthur Samuel, 1959). It explores the study and construction of algorithms that can learn from and make predictions on data[1].

Machine learning is employed in a range of computing tasks where designing and programming explicit algorithms is difficult. Examples of practical applications are numerous and varied: image recognition, medical diagnosis, fraud detection, product recommendations, self-driving cars, etc.

In the case discussed here the machine learning approach involves training an algorithm to identify fuel poor households based on characteristics in the NEED data set.

Figure A1: Traditional vs. Machine Learning programming approach

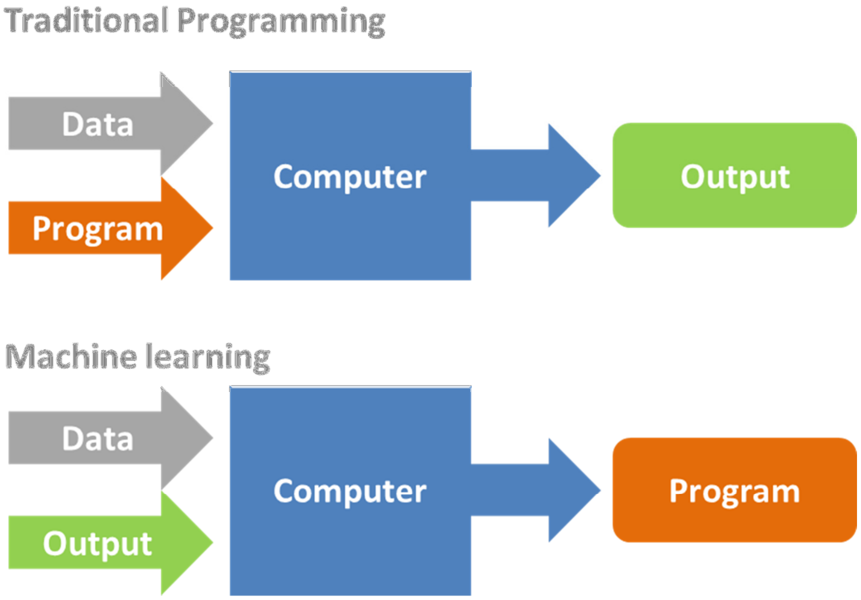


Figure A1 above attempts to illustrate how the task of generating a programme to complete a task (such as predicting the fuel poverty status of a household) is very different when using a traditional programming vs. a machine learning approach. The traditional

[1] https://en.wikipedia.org/wiki/Machine_learning

approach combines data with a user generated program to produce an output. Meanwhile, the machine learning approach takes data and output (i.e. fuel poverty status), and uses a computer algorithm to automatically generate the program that links them.

Data and the modelling approach

In April 2016 a group of experts from across government, industry and academia participated in a fuel poverty targeting workshop to brainstorm both possible relevant sources of data, and potential predictive modelling approaches. Here a number of data sources were identified that could supplement the NEED dataset for this particular task. Following this, a combined dataset capturing the ideas put forward was created, the contents of which is summarised in Table A1 below.

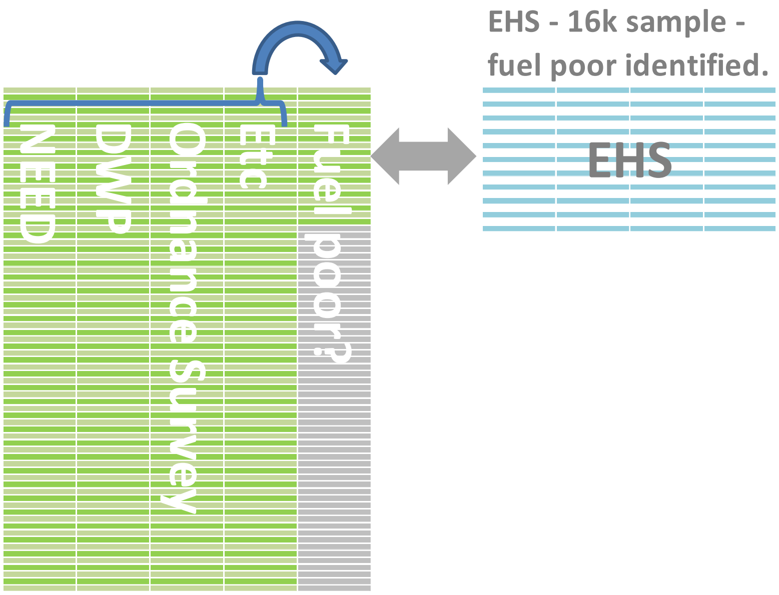
Table A1: Fuel poverty targeting dataset

Source:	Data:
National Energy Efficiency Data-framework	Gas and electricity consumption, energy efficiency installations (e.g. loft insulation)
Experian	Household income, number of adult occupiers, tenure
Department for Work and Pensions	Benefit claimant counts (LSOA level)
Ordnance survey	Building footprint, height and type
English Housing Survey	Fuel poverty status

Modelling approach

Figure A2 below illustrates the modelling approach that was used in this exercise. First a dataset containing relevant details of all households in England was compiled (from the sources noted in Table A1 above). These are illustrated by the green columns in Figure A2. This was then linked to the English Housing Survey (EHS), represented by the blue columns. The EHS contains sufficient information about income and fuel costs to provide an accurate classification of fuel poverty (for the small number of households that were represented in the EHS). This classification (fuel poor/not fuel poor) acts as the 'output' shown in Figure A1. A machine learning algorithm was then trained to establish the relationship between the variables in the full dataset and the fuel poverty status of the household (depicted in Figure A2 by the curved blue arrow). This relationship could then be used to predict the fuel poverty status of the rest of the households in England that were not included in the EHS (as indicated in grey).

Figure A2: The data and modelling approach



Selecting the best predictive algorithm

Selecting the best algorithm with which to make predictions can be a time-consuming manual process – evaluating various possibilities to see which resulted in the highest predictive accuracy. This process is lengthened considerably by the need to experiment with various data transformations and tuning parameter selections, which might impact the algorithms' performance, creating a very large number of transformation/predictor/parameter permutations to evaluate.

When solving a problem such as this, with a large but finite list of possible solutions, time can be saved by employing an automated method. Genetic optimisation is a suitable candidate in this case, an approach that is gaining popularity in the machine learning field to tackle this common issue.

A particular tool[2] was employed that uses this approach to discover the best 'pipeline' (the combination of the choice of: data transformations, predictive algorithm and the associated tuning parameters). The optimisation process starts with a given number of randomly selected pipelines which are then evaluated in terms of their predictive accuracy. Poor performing pipelines are discarded, while the better ones are mutated, before being evaluated again. This process is repeated for a specified number of 'generations', over which the performance of the successful pipeline incrementally improves.

In order to determine the predictive accuracy of each potential pipeline a process known as cross validation is used. This involves splitting the dataset in to two components: a 'training set' and a 'test set'. The training set is used to train the algorithm, which is then used to make predictions of the fuel poverty status of those households in the test set. These predictions are then compared to the actual known fuel poverty status in the test set, in order to determine the proportion of predictions that were correct (i.e. the accuracy). In cross validation this process is repeated, with different randomly drawn test and training sets, in order to safeguard against the possibility that the predictions match the test set by chance. The accuracy of the predictions is then averaged across each of these repetitions to derive a final assessment.

Further algorithm enhancement: Down-sampling

BEIS statistics indicate that roughly ten per cent of households are fuel poor. Consequently, the targeting dataset is dominated by non-fuel poor homes. Disparities of this kind have been shown to distort the 'decision boundary' between classifications, as the larger group exerts a dominating influence. Indeed, an algorithm that classified *all* households as not fuel poor would be correct 90 per cent of the time[3]. In order to avoid this potential pitfall, a technique known as 'down sampling' was utilised. This involves the removal of a random selection of observations of the dominant class (here non-fuel poor) from the training set, in order to reduce its influence. To evaluate the effectiveness of this approach on our fuel poverty classification, a second pipeline optimisation was run on a down-sampled dataset, with equal proportions of fuel poor and non-fuel poor observations. This showed a worthwhile improvement in prediction accuracy compared to the optimal pipeline selected on the full dataset (containing only 10 per cent fuel poor).

[2] TPOT (tree-based pipeline optimisation tool), https://github.com/rhiever/tpot

[3] On the face of it, a prediction accuracy of this order might sound compelling, but for the purposes of locating fuel poor homes, an algorithm of this type would be entirely useless.

Further algorithm enhancement: Grid search

Many machine learning prediction algorithms have parameters that are user defined, i.e. they are not directly learned during the training process. The genetic optimisation used to select the pipeline (described above) attempts to select ideal values for these parameters. In reality, unless the optimisation is allowed to improve over a large number of generations it might be possible to fine tune the parameters further. Grid search is one possible approach to this; a process that allows the exhaustive evaluation of all possible combinations of parameter values, in order to determine which provide the best performance.

Grid search was used in the targeting example discussed here. It was shown to provide a worthwhile increase in predictive accuracy, of a couple of percentage points.

The final algorithm: Random forest

The pipeline optimisation procedure selected a prediction algorithm based on a ‘random forest’. This is a prominent machine learning algorithm which is known to exhibit good performance with data of the type discussed here.

Random forests are an extension of classification trees; a predictive model which maps observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). Rather than relying on a single tree the random forest algorithm creates multiple trees based on a number of repeated samples drawn from the available data and forms a final prediction based on an average of these. The process of forming numerous trees from repeated samples is known as ‘bootstrap aggregation’. Random forests are a special type of bootstrap aggregation where a subset of *m* features are randomly selected from those available, on which to base each tree. When generating a random forest the user has to select two parameters: *n* – the number of trees, and *m* – the number of features used in each tree.

Random forests are known for having a couple of useful qualities:

1. Unlike single classification trees which are susceptible to over-fitting[4], this tendency is averaged out, with a collection of single over-fitting trees cancelling each other out.
2. Using a sub-set of features for each tree allows the importance of each potential feature to be examined in isolation of other features that might mask its potential influence[5].

[4] Over-fitting is a situation where the algorithm produces predictions that fit well in a manner that is highly specific to the training set, with a consequence that it makes poor predictions of the test set.

Results

Prediction results of machine learning algorithms are often presented in a ‘confusion matrix’. This format is used below in Table A2, which shows the number of households in the test set that are predicted to be fuel poor (or not), and how many of these positive or negative predictions were correct. The four cells in the table therefore represent correct positives (CP), false positives (FP), correct negatives (CN) and false negatives (FN).

Table A2: Confusion matrix of fuel poverty predictions

		Actual	
		Fuel poor	Not fuel poor
Predicted	Fuel poor	CP: 266	FP: 865
	Not fuel poor	FN: 30	CN: 1531

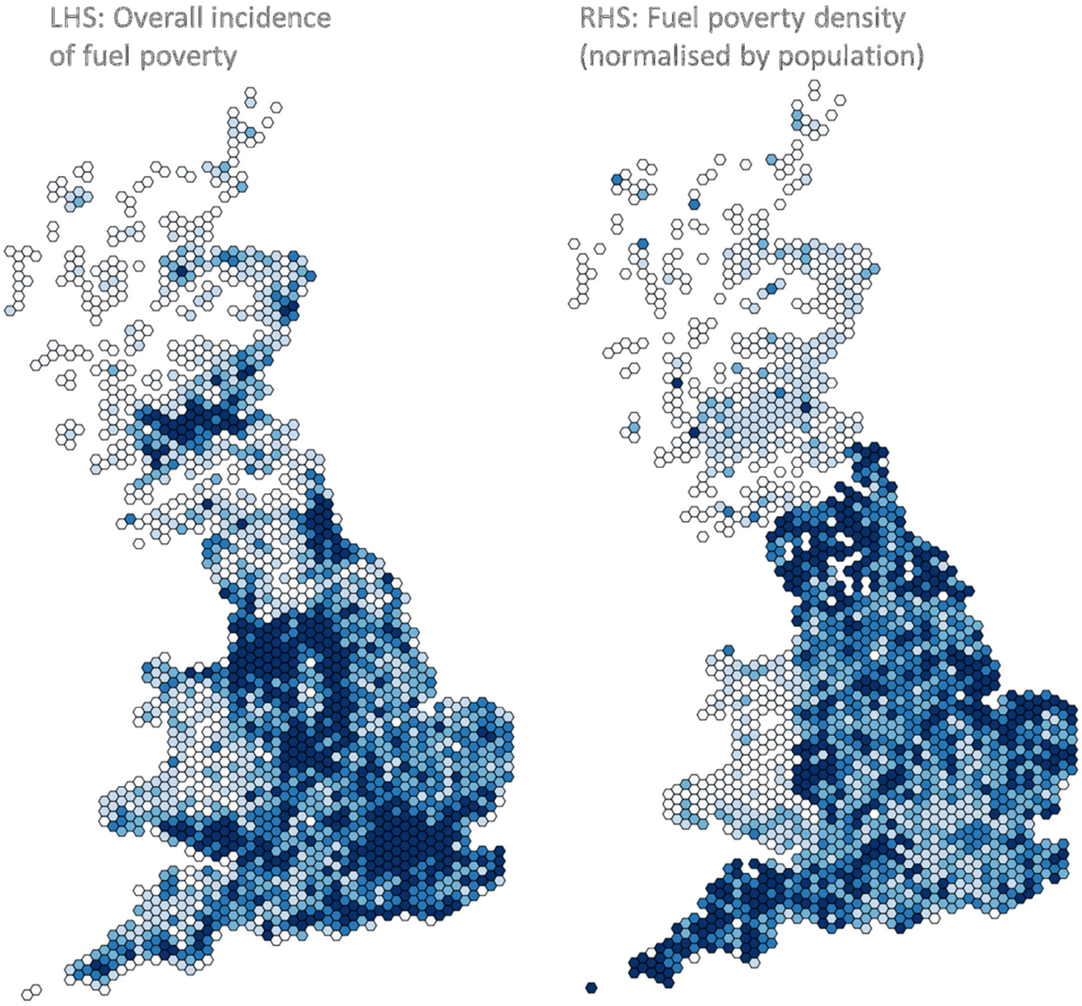
The proportion of households predicted to be fuel poor is higher than the true proportion, i.e. the selected algorithm has a tendency to make false positive predictions. A corollary of this is that among those households that are predicted to be fuel poor only around one in four actually are. Although this is not ideal, it does represent an improvement over a random-guess baseline, where only one in ten fuel poor predictions would be correct. The not fuel poor predictions a more accurate (98 per cent correct). This would help ensure that any potential targeting mechanism based on the algorithm would overlook only a small proportion of actual fuel poor households (10 per cent).

As the algorithm makes predictions at the household level, these can be summarised in geospatial plots. The left-hand map in Figure A3 below shows the regional incidence of the fuel poverty predictions (darker blue denotes more fuel poverty). Unsurprisingly, the number of fuel poor households tends to be higher in densely populated areas, simply because the number of households overall is higher. To account for this, the right-hand map is normalised by the number of households, giving a representation of the ratio of fuel poverty. In addition to England, an attempt was made to use the algorithm to make fuel

[5] For this reason random forests are often run prior to building regression models, as a means of identifying the best set of predictors to include in the model.

poverty predictions for both Scotland and Wales. Given that the training set only contained data for England, this was perhaps slightly ambitious; consequently Figure A3 shows that the algorithm has failed to make a realistic number of positive predictions in these two countries. This suggests that fuel poor households across Great Britain are not entirely homogenous, and that training data for all three countries would be required to make predictions for them.

Figure A3: Regional summary of fuel poverty predictions



Comparison with similar work

This work has many parallels with BEIS' regional fuel poverty estimates[6]. As shown above, it can be used to produce a regional overview of fuel poverty. A side-by-side comparison of these two analyses does indeed reveal some broad consistencies. There are however two distinctions of note. Firstly, whereas the regional estimates are primarily concerned with establishing the absolute level of fuel poverty in a given area, the targeting algorithm seeks to maximise the ratio of correct household level predictions, which is a slightly different objective. As shown in Table A2 above, this ratio happens to be maximised where a number of false positives occurs, which might lead to a distortion in the regional pattern if for some reasons these false positives didn't occur randomly across the country. Secondly, the targeting approach takes a bottom-up household level approach, while the regional estimates are made at an aggregated level.

[6] https://www.gov.uk/government/collections/fuel-poverty-sub-regional-statistics