BENV0096: MSc ESDA Dissertation

"Predicting Heating Operation from Gas Meter Data"

by

TYGR3

September 2022

Paper submitted in part fulfilment of the
Degree of Master of Energy Systems and Data Analytics

Energy Institute
University College London

Word count: 9400

Acknowledgements

I would like to thank my supervisors, Despina Manouseli and Jenny Crawley, for many invaluable insights, continuous support and encouragement through the course of this dissertation project. I am sincerely grateful for the time that you spared sharing your knowledge and providing meticulous feedback. I would also like to thank my wife, my dog and my family for their constant love and support.

Abstract

Understanding heating operation in homes is crucial to transition to alternative energy sources. This study proposes a methodology that uses indoor temperatures to generate heating state labels, which are then used as ground truth to train classification models identifying heating operation from gas consumption and external temperature.

The heating state labels are generated from living room temperatures and average indoor temperatures from a set of homes, which are then validated against heating flow temperatures. The labels generated from average indoor temperatures are found to be more robust and are used as ground truth labels to test and compare the performance of four classification algorithms: Logistic Regression, Decision Trees, Random Forest, and XGBoost.

The models using the four algorithms are trained on gas consumption, rolling statistics and lagged variables extracted from gas consumption, date-time features and external temperatures. Rolling statistics are found to be the most significant in modelling, while the majority of date-time features were not found important. The heating labels are found to be heavily imbalanced and the performance of the models significantly improves after oversampling the training data. Logistic Regression and Decision Trees models are found to be the best performing, identifying heating operations with an F1-scores of 65% and 66% respectively.

Table of Contents

List of Figures

List of Tables

1 Introduction & Project Framework

1.1 Introduction

In 2019 the building sector in the UK accounted for 23% of the total greenhouse emissions, with 77% of the direct CO2 emissions resulting primarily from the use of fossil fuels for heating homes [1]. Hence, emission reduction in buildings is crucial to meeting the UK' s ambitious commitment to reach net zero by 2050 [2]. Moreover, the cost-effectiveness of energy conservation measures, coupled with the large proportion of energy consumed in buildings, creates great opportunities for energy saving and decarbonisation [3]. To benefit from these opportunities a mix of policies is implemented that includes requirements and incentivisation of behaviour change, energy efficiency improvements and switching to low-carbon heat [4]. Nevertheless, carbon emissions reduction in residential buildings is identified as one of the biggest policy challenges for the government [5].

Often, the policy mix includes switching high carbon fossil fuels to a system heavily reliant on heat pumps. Adoption of the renewable energy sources in the UK is leading to a progressively decarbonised electricity supply [6]. This makes the electrification of heat with a system heavily reliant on heat pumps powered by low-carbon electricity an attractive solution for policymakers for decarbonising heating, cooling, and hot water in residential buildings [7]. However, this would increase the electricity demand and reliance on highly variable renewable energy sources, with a 30% shift to heat pumps in domestic heating estimated to increase the total UK electricity demand by 25% [8]. The potential negative impacts from this, such as the increased peak demand, create a need for reducing the energy demand and increasing flexibility [9].

As heating and hot water in buildings account for 40% of energy consumption in the UK [10], managing their demand could help increase the flexibility of the whole electricity system. This can be achieved by exploiting the energy storage potential in buildings to time shift heating or cooling [11] and allowing

buildings to act as prosumers, both consuming and producing the energy [12]. However, the housing stock in the UK is among the worst compared with other European countries [13], making it difficult to achieve this flexibility without jeopardizing the occupants' comfort [14].

Therefore, it is essential to benchmark the energy consumption in buildings through modelling to make policies informed by the energy use in the housing stock. In the past, this has been done predominantly using steady-state models, such as the Building Research Establishment Domestic Energy Model (BREDEM), which is widely used in the UK as a foundation for the primary energy efficiency assessment mechanism as well as many other UK building stock models, such as DeCARB, CDEM, BREHOMES, UKDCM and The Cambridge Housing Model [15]. However, the steady-state models assume a single temperature and heating pattern, standardising occupant influences, which can be unreliable and not sufficiently based on robust data [16]. Many studies in the past have shown discrepancies when comparing model-based predictions to actual energy consumption or heating demand [17]–[19]. These discrepancies eventually result in a performance gap between the savings achieved by the implementation of policies informed by models such as BREDEM and the theoretical predictions [20]. To address the performance gap, it is necessary to implement methods that would allow us to understand the reasons behind the gap, weather-related to human or physical factors, assessing the performance of existing buildings at scale [21].

An alternative measure of energy use in buildings to the steady state models are dynamic models [22]. Dynamic models provide accurate and detailed outputs that can be applied intelligently for demand-side management, optimal control, and characterisation of fabric performance [23]–[25]. However, dynamic energy models of buildings require high-frequency time series of the heating output or the energy used for space heating. As such, these models usually require increased instrumentation and intrusive heating experiments and their implementation on large scale with minimal costs and disruption of occupants remains a challenge. Although the recent deployment of smart meters has allowed easy and unintrusive collection of energy

consumption data required for the dynamic models, the smart readings record the total fuel usage of the buildings. In the case of natural gas, this includes cooking and hot water, and the energy used for the space heating has to be separated. This challenge is often referred to as disaggregation or single-channel source separation [26] and has been studied extensively in similar cases for electricity meters, often using dynamic pattern recognition [27]. However, there are only a few studies that consider the gas meter data, that require disaggregation to separate the sums, instead of the high-resolution time series [21].

1.2 Research Question and Contributions

This study aims to develop a robust methodology for identifying active space heating periods from half-hourly gas smart meter data in the residential sector. For this purpose, a novel IDEAL dataset is employed, including internal temperatures and central boiler heating flow temperatures alongside the gas meter readings [28]. Should the active space heating periods be accurately labelled using the internal temperatures and heating flow temperatures, the labels could be further employed as ground truth in classification models. This would offer a unique opportunity to treat the disaggregation of the gas meter data into heating and non-heating periods as a classification problem and compare the available classifiers to develop a robust and accurate methodology. This thesis will attempt to answer the following research questions:

- What can be inferred from the internal temperatures, heating flow temperatures and gas meter data in the IDEAL dataset?

- Can active space heating periods be accurately labelled using internal temperatures? What internal temperatures produce the most accurate labels?

- Can active space heating periods be accurately classified in unseen data by the classifier model trained on half-hourly gas meter data with daily external temperatures and validated using the generated labels? Which classification algorithm performs the best for this task?

- How does the performance compare between the models trained on unbalanced and models trained on balanced space heating labels?

2 Literature review

2.1 Energy consumption disaggregation

There is a large body of literature that studies single-point metered energy consumption disaggregation by its end uses. Otherwise known as Non-Intrusive Load Monitoring (NILM), it has been first developed by MIT in the 1980s and was used to successfully detect on-off appliances and finite state machines [29]. The established version of the original methodology used the known consumption power loads to associate with it the clusters of changes in steady-state power draw levels in the signature space of real/reactive power [30]. Other earlier studies have used different measures relevant to electric loads, such as voltage and current loads [31]. While in the case of nonlinear appliances, high-frequency sampling is used for transient state analysis, which includes spectral analysis and wavelet transformation [32], [33].

Overall, the majority of the NILM studies use supervised machine learning methods, which are predominantly either pattern recognition or optimisation-based approaches [27]. However, the optimisation approaches used in the past, which included integer programming and genetic algorithm, are computationally expensive and unable to discern appliances with overlapping or similar load signatures [34]–[36].

One of the early studies used the decision rules specific to the appliances, alongside the changes in real power, to detect large appliances with an arbitrary score system that constitutes the pattern recognition approach [37]. The pattern recognition approach is more popular in modern studies, which often use smart meter-like datasets of a single measured quantity, such as real power, at intervals of 1 second or more [27]. This approach finds the most matching sequence of appliance switching using the appliance signatures collected previously by experimentation [36].

Although pattern recognition can be used in a large variety of problems, including water metering [38], due to the differences in switching behaviour

of various systems the developed solutions can only be used for the intended purposes. This includes the gas readings data that has only a few available related studies. There are already existing statistical approaches for gas data disaggregation [39], but they are used for disaggregating total volumes over a time period, while dynamic models require usage in a single time step.

One study uses heat flow meter data with a 10-minute resolution, differentiating the space heating load from the water heating load using the comparison between periods of typical operation and a period of space heating operation with no occupancy [40]. Despite the study using the total heat load instead of the gas data, the objective here is the same. In this study, a non-parametric estimator is designed to filter out the spikes of water heating, which takes the form of short and large spikes, unlike space heating which is observed to be continuous, slow-changing, and low relative magnitude. However, these assumptions are not true for gas boilers, with spikes often associated with missed heating periods.

Another paper uses dynamic pattern matching of gas consumption data to separate the hot water, cooking and heating [21]. The space heating is observed to have a consistent signature, allowing pattern recognition to be an effective method for this problem. The study compares the space heating signatures to the gas usage profile in all-time windows of non-transient activity and uses a similarity metric to select more similar windows.

2.2 Inferring heating operation from indirect data

A less explored alternative approach uses indirect sensor data collected in homes, such as room or radiator temperatures, to infer heating operation.

A study by Shipworth et al. [18] infers heating duration and thermostat setting in 358 homes. Room temperatures collected every 45 minutes are used in the study, assuming that the rise in indoor air temperature during winter months is associated with space heating. The average thermostat setting was inferred to be 21.1°C, which supports the heating demand assumptions used by BREDEM. No correlation was found between the estimated and reported values, which is attributed to social desirability, lack of technical understanding and frequent adjustments. The heating duration was estimated to be 8.3h on weekdays and 8.4h on weekends, which is in line with the BREDEM assumption for weekdays but differs from the weekend heating assumption of 16h. Both the thermostat setting and heating duration have shown large variations between homes in the study.

In a study by Huebner et al. [15] mean heating duration was calculated from a sample of 248 homes from the same dataset. The method assumes the change of state if the cumulative temperature change for a number of steps, until the sign of change inverses, is larger than 0.75 °C. The average heating duration is estimated to be 10h per day. This supports Shipworth et al. [18] finding that the heating duration is similar between the weekends and the weekdays, although different in value due to the differences in estimation method. Compared to the BREDEM model assumptions the estimated heating duration is smaller by 7h on a weekly basis, with a substantial portion of sampled homes found to be outside the heating periods assumed by the model. The study reported significant variations in heating duration between homes, which also supports the findings of Shipworth et al. [18]. The similar heating duration between the weekends and the weekday and the variations between the sampled homes are also found in the author's previous paper [17], which was focused on comparing the measured living room temperatures with heating periods assumed in BREDEM (Figure 1).

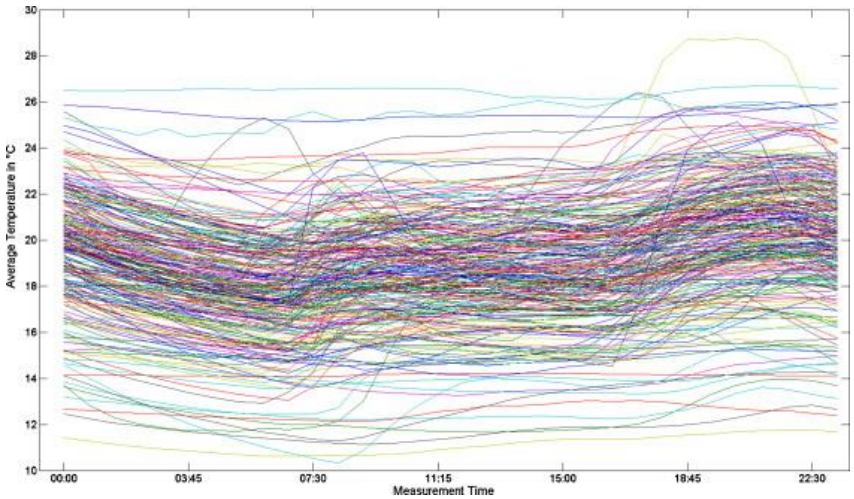


Figure 1. Average weekday winter temperatures for 248 homes are indicated by individual lines [17].

Kane et al. [41] compare methods used by Shipworth et al. [18] and Huebner et al. [15] as well as three other approaches to generating heating state labels. One of the assessed methods considered thermostatic control, labelling heating as off if it remains constant for three time steps, using hourly data [42]. These methods, as well as another method introduced in the study that used gas consumption data, were found less reliable, especially in the summer months. The method that was found the most reliable used direct measurements of radiator temperature, although such data was noted to be hard and costly to obtain. The study also finds the mean thermostat setting to be range between 18 °C and 22 °C and subject to significant uncertainty.

Another study by Kane et al. [43] uses hourly room temperature data collected from 249 homes between 1st December and 28th February to calculate nine heating practice metrics. The metrics include timer settings and heating duration and estimate the mean heating duration of 12.6h, which is longer compared to other studies in the literature. The mean thermostat setting was estimated to be 20.9 °C, which supports findings from Shipworth et al. [18].

A recent paper by Crawley et al. [44] uses a combination of gas and room temperature data from the IDEAL dataset to assign heating state labels, which are then used to develop an empirical energy demand flexibility metric. The method assumes that the heating is "on" if the use of gas corresponds with an increase in room temperature, either immediately or with a time offset.

The IDEAL dataset is also used in the study by Berliner et al. [45] to extend on methodology evaluated in [41] and [17] and infer the use of space heating from room temperature and humidity. The method uses radiator temperatures at 10-minute granularity to capture periods when the radiator is hotter than the room temperature by 5 °C or more. These periods were then used as ground truth labels in the classification model predicting if space heating is on or off at each time period with room temperature and humidity as inputs. A deep, dilated convolutional network, which is a form of artificial neural network (ANN), is used for classification, achieving precision and recall of 0.73 and 0.75 respectively. The model performance was reported lower during the summer months.

3 Methodology

This chapter includes 5 sections that describe the methodology. Section 3.1 will describe the data used in this study. Section 3.2 will describe the initial processing steps that were applied to the data. Section 3.3 will describe the creation of heating state ground truth labels. Section 3.4 will describe the classification models used in this study and what data preparation has been performed prior to the classification. Section 3.5 will describe the classification performance metrics used in this study.

3.1 Dataset description

The data used in the present study was sourced from the IDEAL household energy dataset [28]. The IDEAL dataset contains time series of high-resolution gas and electricity consumption data, linked to a wide range of contextual sensor and survey variables. The dataset was collected over various periods between August 2016 and June 2018 from 255 participating households located in Edinburgh and the surrounding regions of the Lothians and south Fife, in Scotland, UK (Figure 2). Periods of data collected from different homes ranged from 55 to 673 days, with a mean of 286 days.

The sample selected for this study included homes which had data covering the period from January 2017 to June 2018. This was done with the intention to use the full year of 2017 for training and January to June 2018 for validation in the classification models that are described in section 3.4. Full-year of training data would allow to account for seasonal variability that is inherent to energy consumption time series [46].

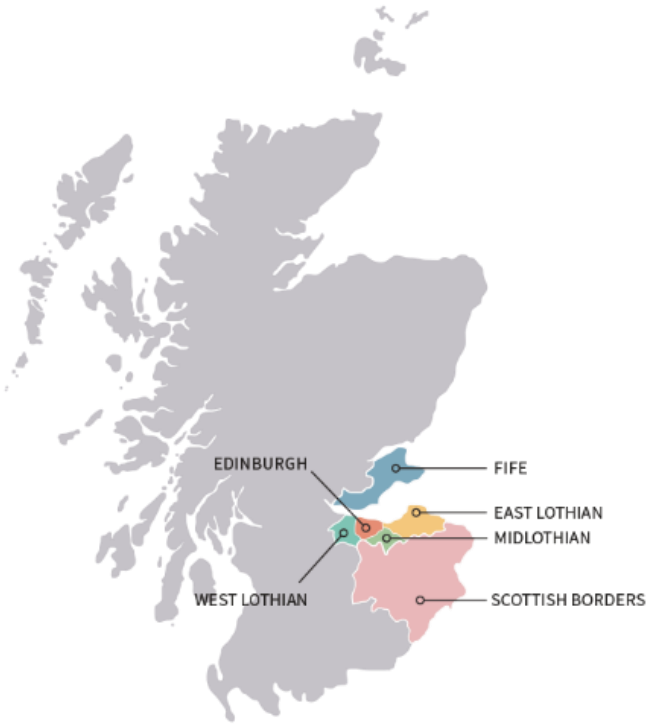


Figure 2. Map of Scotland showing Edinburgh and the surrounding regions [47].

Out of the total of 255 participating households, only 10 contained data from January 2017 to June 2018, of which two were omitted due to extended periods of missing values, leaving a total of 8 homes (Table 1). Although all of the selected homes are located in large urban areas of Edinburgh, they make up a heterogeneous group in terms of physical properties and their residents.

Characteristics of homes and their residents can significantly affect consumption patterns. Flats and modern buildings are more likely to be better insulated, which leads to smaller demand [48]. In contrast, a larger number of residents, higher income bands, and a higher number of occupied days are often associated with higher energy demand [49]. Hence, a spread of characteristics in the sample is important to account for this variability.

Table 1. Characteristics of homes selected for this study and their residents.

Home ID	Home type	Built era	Entry floor	Outdoor space	Income band	Occupied		Resident number	Gender	Age	Work Status	Work hours (weekly)	Education
						Days	Nights						
62	Flat	1850-1899	2nd	Yes - shared	£43,200 to £48,599	1	7	2	Female	30-34	Paid work	31-40	Degree level or equivalent
									Male	30-34	Paid work	31-40	Degree level or equivalent
63	House / bungalow	1919-1930	Ground	Yes - private	£54,000 to £65,999	2	7	2	Female	35-39	Paid work	41-50	Degree level or equivalent
									Male	40-44	Paid work	31-40	Degree level or equivalent
64	Flat	Before 1850	1st	Yes - shared	£66,000 to £77,999	6	7	4	Male	0-4	N/A	0	N/A
									Female	0-4	N/A	0	N/A
									Male	35-39	Paid work	41-50	Degree level or equivalent
									Female	35-39	Paid work	31-40	Degree level or equivalent
66	House / bungalow	1965-1980	Ground	Yes - private	£66,000 to £77,999	4	7	3	Male	0-4	N/A	0	N/A
									Female	0-4	N/A	0	N/A
									Female	35-39	Paid work	21-30	PhD
									Male	30-34	Paid work	31-40	PhD
67	Flat	1945-1964	Ground	Yes - private	£27,000 to £32,399	2	7	2	Female	50-54	Paid work	41-50	O Level or GCSE equivalent
									Male	50-54	Paid work	41-50	O Level or GCSE equivalent
68	House / bungalow	1945-1964	Ground	Yes - private	£48,600 to £53,999	4	7	1	Female	50-54	Paid work	21-30	Higher educational below degree level
70	Flat	2002 or later	4th	Yes - shared	less than £10,800	2	7	1	Female	25-29	Student	1-10	Degree level or equivalent
73	Flat	1850-1899	Ground	Yes - shared	Missing	7	7	3	Female	10-14	Student	0	No formal qualifications
									Female	15-19	Student	0	GCSE grade D-G or equivalent
									Male	45-49	Self-employed	1-10	A-Levels or Highers

Three data types collected from the selected homes were used in this study (Figure 3):

1. Pulse-level gas consumption data in Watt hours (Wh) at 1-second intervals.
2. Indoor temperature for each room in degrees Celsius (°C) at 12-second intervals.
3. Heating flow temperature (or central heating flow temperature) in degrees Celsius (°C) at 12-second intervals.

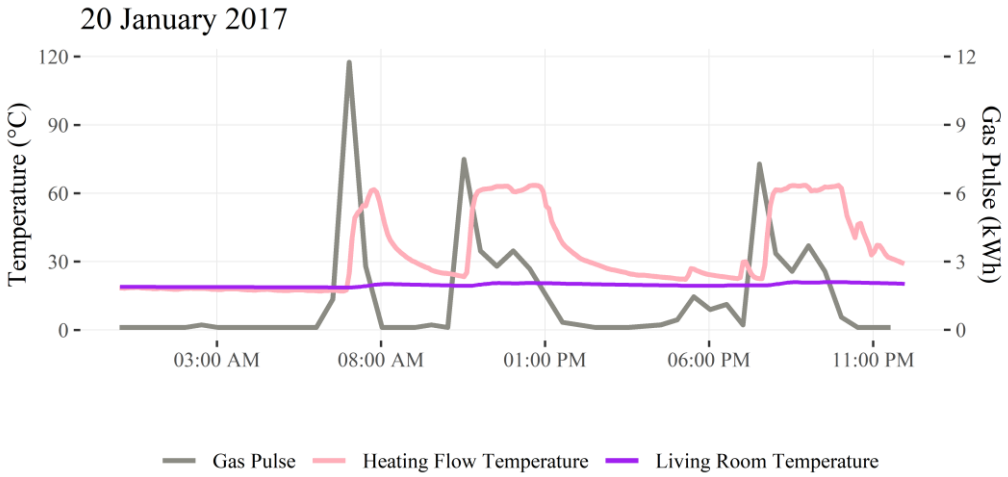


Figure 3. Example day showing three datatypes collected from Home 62.

Moreover, the daily external temperature in degrees Celsius included in the IDEAL dataset was also used.

3.2 Initial data processing

Prior to analysis and modelling the gas consumption, indoor temperature and heating flow temperature data was cleaned and downsampled.

The pulse-level gas data was downsampled to 30 minutes, as to replicate the standard granularity of consumption data collected in gas smart meters. The sensors in the IDEAL study recorded consumption as a number of pulses of 112 Wh, with any usage below that value not recorded. Therefore, a sum total of consumed gas was taken for each downsampled interval, while the missing data was interpreted as periods of low to no usage and imputed with a single pulse value of 112 Wh. Gas smart meter data is widely collected in 30-minute intervals in the UK, with over 40% of the gas meter reported smart in smart mode at the end of March 2022 (Figure 4). Hence, conducting this study on gas smart meter-like data will allow the developed methodology to be applicable to more than 9 million homes in the UK.

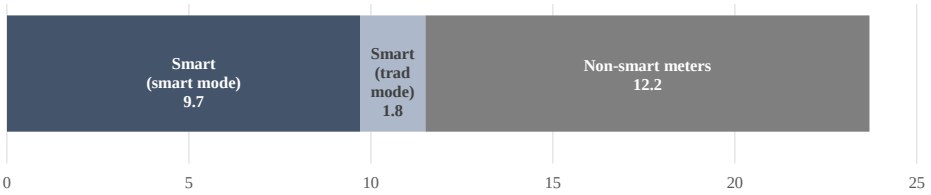


Figure 4. The number of smart and non-smart gas meters in operation in Q1 2022 in the UK, millions [50].

The indoor temperature and heating flow data was downsampled to 5-minute intervals, taking the mean of the values in each interval. This was done to reduce the volatility in data and allow additional robustness in heating state label generation algorithms described in section 3.3. Missing temperature data were imputed using the simple and effective Last Observation Carried Forward (LOCF) method, which involves replacing the missing values with the most recent previously observed value. LCOF is widely used and has been reported to perform robustly on a variety of data types, including indoor temperature, outperforming other imputation methods such as imputation using mean or most frequent value [51].

Figure 5 shows the gas and temperature data before and after the transformation.

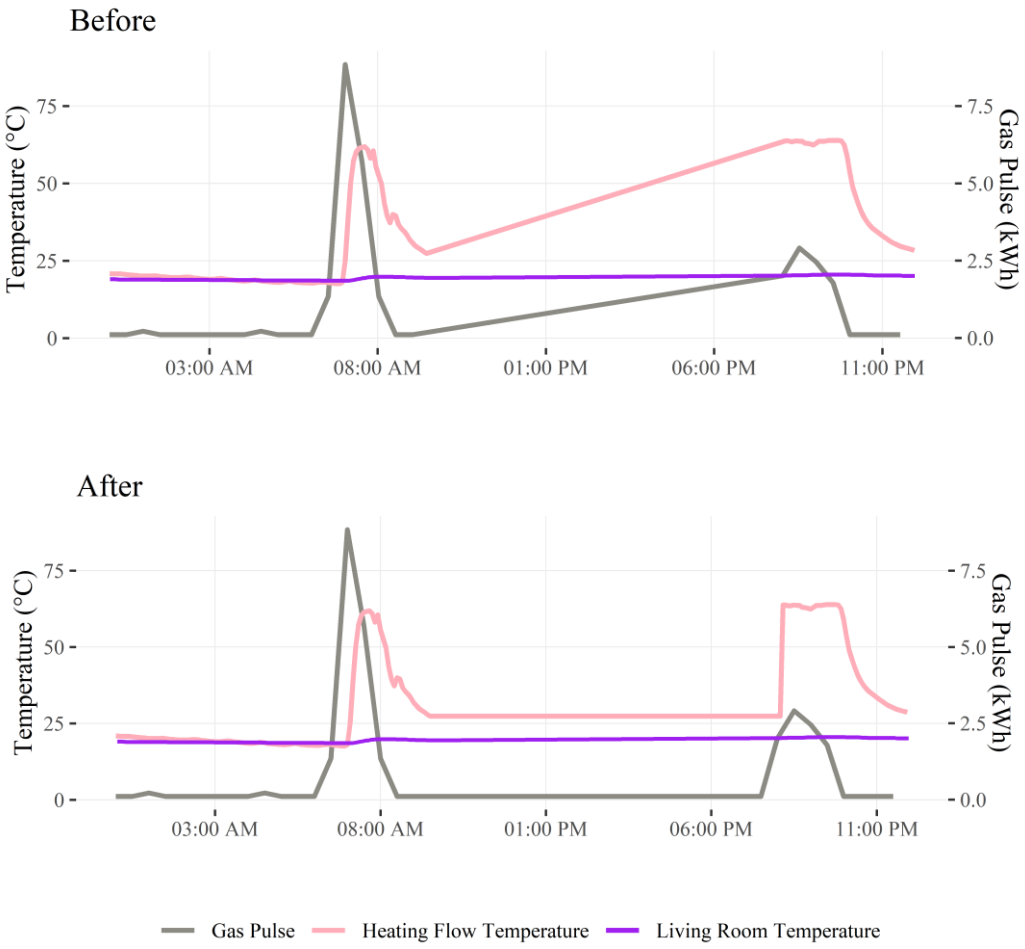


Figure 5. Example of missing data before and after the imputation on 2 February 2017 from Home 62.

Living rooms are conventionally heated most frequently in dwellings, and as such, are often employed in energy consumption analysis [41]. However, considering the availability of temperature data for every room in the house in the IDEAL dataset, indoor temperatures averaged for the whole house are used alongside the living room temperatures in this study (Figure 6).

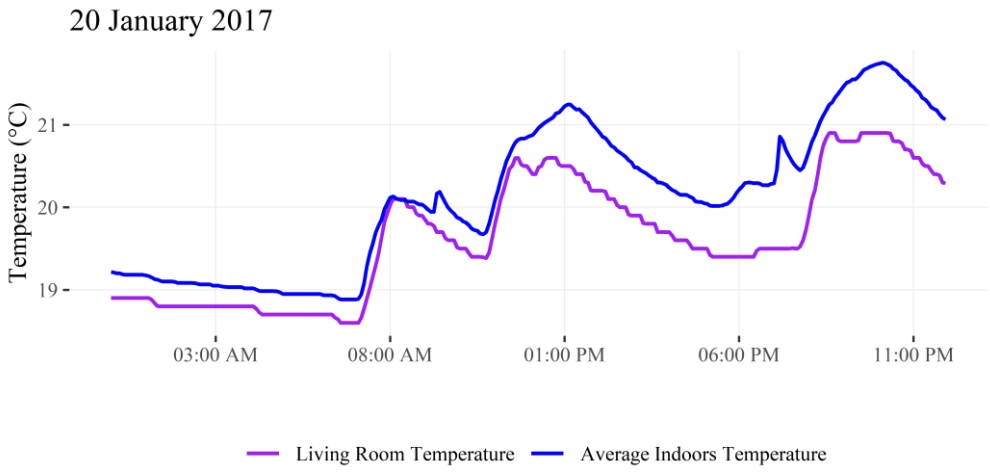


Figure 6. Example day showing living room and average indoor temperatures collected from Home 62.

3.3 Ground truth labels

This section describes the generation of ground truth labels that are required to treat gas consumption disaggregation as a classification problem. In this work, three methods were used for generating the ground truth labels: *M1*, *M2*, and *M3*. The methods involve the use of rule-based algorithms, which are based on conditional statements that the input data is measured against.

M1 and *M2* used the same rule-based algorithm, however, living room temperatures were taken as input in *M1* and average indoor temperatures were taken as an input in *M2*. The algorithm was developed following the method described in [15], however using temperatures with 5-minute granularity. Examples illustrating each step of the algorithm are provided in Table 2.

First, a vector of difference values is calculated from the temperature difference between points t_{n+1} and t_n (*Step 1*). By locating the point in this vector where the difference values have shifted from positive to negative or vice versa, rising temperature sequences were identified (*Step 2*). Zero difference values were deemed to be negative since we are only interested in periods of positive increment. Then, the magnitude of change was determined by taking the total sum of all difference values in each identified sequence (*Step 3*). Finally, the sequence is labelled as *"On"*, if the magnitude of change is larger than the set criterion of *0.75 °C* (*Step 4*), which reflects the average reported hysteresis of the thermostat [15].

Table 2. Examples Illustrating the rule-based algorithm used in M1 and M2.

	Temperature values	17.8	17.4	17.2	17.3	17.5	17.2	17.5	18.2	18.7
Step 1:	Temperature differences		-0.4	-0.2	0.1	0.2	-0.3	0.3	0.7	0.5
Step 2:	Change sequences	1	1	1	2	2	3	4	4	4
Step 3:	Magnitude of change			-0.6		0.3	-0.3			1.5
Step 4:	Heating state labels	Off	Off	Off	Off	Off	Off	On	On	On

The third method, *M3*, takes heating flow temperatures as input. Heating flow temperatures are measured from the boiler flow pipe, reporting the temperature of the water that leaves the boiler before it circulates the heating system. This ensures that the data is not polluted by interferences from cooking, external heating sources and other occupant behaviour that are often present in indoor temperatures [41]. Moreover, the rise of temperatures during active heating is more prevalent in heating flow temperatures than in indoor temperatures. Therefore, the outputs of the third method were used for validation of *M1* and *M2*.

The algorithm used in the *M3* follows a simple rule, labelling the heating system as active when the temperature of the heating flow exceeds a threshold. Considering that the heating flow temperature of a standard boiler in the UK is 55°C [52], the threshold value is taken as *45°C*, leaving a 10°C gap to account for differences in heating demand in homes, which can be influenced by heating behaviours and preferences of the occupants as well as physical properties of the dwellings.

The statement labels were then translated into binary labels where *"Off" = 0* and *"On" = 1*, which were then downsampled to 30-minute intervals. The maximum label value was taken for each downsampled interval. Following that, to reduce the imbalance in the data was filtered to only include the heating months, when the heating events happen more often. Finally, for both filtered and unfiltered data, the performance of the labels generated with *M1* and *M2* was measured against the labels generated with *M3* using a set of metrics to determine which labels to use for classification.

3.4 Classification

Using the ground truth labels identified in the previous section, a classification model was developed to predict heating operation. Indoor temperatures and heating flow temperatures are not used in the classification, only employing the half-hourly gas meter (Wh) and daily external temperature data (°C), which is widely accessible [41]. This would allow scaling the identification of heating operations beyond homes that have participated in intrusive studies.

3.4.1 Data preparation and feature engineering

To avoid data leakage, prior to any feature engineering the dataset was split into test, validation, and train. Homes 63 and 67 were chosen for testing, while in the remaining homes the period of 10.01.2017-31.12.2017 was used for training and 01.01.2018-01.06.2018 for validation. The same data preparation and feature engineering were performed on all three parts of the dataset.

Firstly, the outliers for gas consumption and external temperatures were removed using a commonly used rule of thumb that suggest three times the standard deviation from the mean of the normal distribution as a threshold [53]. Following that, the time-series dataset was transformed by removing the temporal ordering of individual input examples and extracting a set of features that provide the temporal information, including date-time variables, lagged variables and rolling statistics [54]–[57].

Date-time variables

Following date-time variables are extracted to explain the daily and seasonal patterns in data [54]:

1. *hr* – hour of the day, 24 unique values;
2. *yday* – day of the year, 365 unique values;
3. *mday* – day of the month, 31 unique values;
4. *wday* – day of the week, 7 unique values;
5. *wend* – binary variable labelling the weekends;
6. *week* – week of the year, 52 unique values;
7. *mnth* – month of the year, 12 unique values;
8. *qrtr* – quarter of the year; 4 unique values;

Although some of the date-time variables are likely to be redundant, there is no need for feature selection due to a large number of observations.

Lagged variables

Lagged variables provide the values of the variable at previous time steps [55]. Lagged variables within x_{t-z} are considered for any value of gas consumption x_t. Upon examination of autocorrelation and partial autocorrelation of the gas consumption (Figure 7), it was determined that $z = 5$ would be optimal and lagged variables *l1*, *l2*, *l3*, *l4*, and *l5* were extracted.

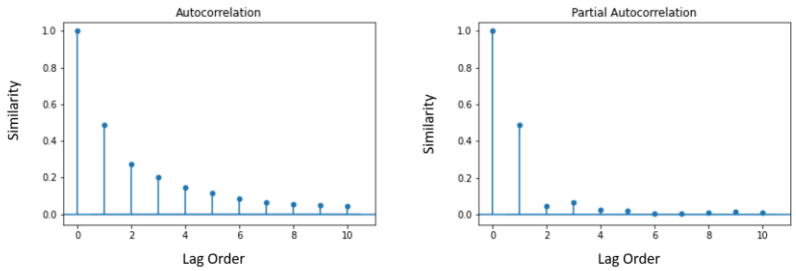


Figure 7. Autocorrelation and partial autocorrelation between the gas consumption and its lagged values.

Rolling statistics

Six rolling statistics have been extracted from gas consumption values with a window of six time steps [56], [57]:

1. *rmean* – rolling mean, which is a mathematical average of values in a set window.

2. *rmedian* – rolling median, which is the middlemost observation of values in a set window arranged by magnitude.

3. *rslope* – rolling slope, which is a measure of how much the values in a set window change with each time step.

4. *rstd* – rolling standard deviation, which is a measure of how spread out the values are from the mean in a set window, calculated by $\sigma = \sqrt{\frac{\boldsymbol{X}-\overline{X}}{n}}$, where n is the size of the sample $\boldsymbol{X} = X_1, X_2, \dots, X_n$ with mean $\overline{X}$.

5. *rskew* – rolling skewness, which is a measure of the lack of symmetry of values in a set window, calculated as $skew = \frac{(\boldsymbol{X}-\overline{X})^3/n}{\sigma^3}$, where n is the size of the sample $\boldsymbol{X} = X_1, X_2, \dots, X_n$ with mean $\overline{X}$ and standard deviation σ.

6. *rkurt* – rolling kurtosis, which is a measure of whether the values in a set window are light-tailed or heavy-tailed, it is calculated as $kurt = \frac{(\boldsymbol{X}-\overline{X})^4/n}{\sigma^4}$, where n is the size of the sample $\boldsymbol{X} = X_1, X_2, \dots, X_n$ with mean $\overline{X}$ and standard deviation σ.

Once all the features were extracted, the categorical variables were transformed into factors, leaving only numerical and factor data types (Figure 8) and the dataset was filtered to only include the heating months (October to April).

Figure 8. Datatypes of predictors and the predictand in the dataset.

Finally, to tackle the bias created by the imbalance of heating state labels the minority class was oversampled in the training dataset (Figure 9).

Unbalanced

On	Off
7299	51141

Balanced

On	Off
51141	51141

Figure 9. The number of labels in each class in the training dataset before and after oversampling.

3.4.2 Classification Algorithms

Four different classification algorithms were compared to find the best performing classifier: Logistic Regression, Decision Tree, Random Forest and XGBoost.

Logistic Regression

Logistic regression is a statistical algorithm that is often used in supervised learning. It has been previously used in literature for energy demand forecasting [58] and the classification of electricity consumption [59], achieving high performance. As a classification algorithm, logistic regression predicts the probability of a target variable, learning a linear relationship from the data and creating non-linearity characteristics through Sigmoid functions. Mathematically logistic regression can be expressed as:

$$\log\left(\frac{p}{1-p}\right) = y$$

Where p is the probability of success with a value between 0 and 1, y is the linear model and $p/(1-p)$ is an odd ratio. For the implementation in this study, the output of the hypothesis function is interpreted as positive if it is ≥ 0.5 and negative otherwise.

Decision Tree

Decision trees are a supervised machine learning algorithm, that continuously splits the data according to a parameter and is explained by decision nodes and leaves. In classification, the tree is represented by yes/no conjunctions of features that lead to the class labels, with splits decided by Gini impurity. The Gini impurity can be fined as:

$$Gini(D) = 1 - \sum_{i=1}^{k} p_i^2$$

Where D is a dataset that contains samples from k classes and p_i the probability of a sample belonging to class i at a given node.

Decision trees have been commonly used in data mining [60]. They have also been used in building energy demand modelling, achieving accurate predictions and demonstrating high interpretability [61]. The major advantage of the decision tree algorithm is that it produces models that represent interpretable rules or logic statements and provides information on the significance of the factors for prediction or classification. However, decision trees are susceptible to noisy data and do not perform as well on nonlinear data as neural networks [62]. Instead, the algorithm is better suited for predicting categorical outcomes or time series data with the clear trend and sequential patterns.

Random Forest

Random forest is a classification algorithm that consists of an ensemble of decision trees. The algorithm builds and combines multiple decision trees to increase accuracy in predictions, calculating the prediction by summing the prediction of the ensemble. In previous studies, random forest has shown high accuracy and performance in classification [63]. The algorithm has also been used to predict hourly electricity usage in educational buildings [64] and classify domestic smart meter data [65] with high accuracy.

XGBoost

XGBoost, which stands for Extreme Gradient Boosting, is a commonly used machine learning method. XGBoost is a scalable distributed implementation of gradient boosting decision trees (GBDT) that provides parallel tree boosting. Essentially, the algorithm is creating decision trees in sequence with weights that can be tuned and inputted into a decision tree. In the literature, it has been used for energy prediction and feature importance evaluation achieving high accuracy [66]–[68].

3.5 Classification performance metrics

Performance metrics are a key aspect of building a classification model. Determining the model performance allows us to optimise the classifier parameters and compare the final results, selecting the optimal algorithm. However, as the classifier can perform well in certain classes while obtaining poor scores in others, several metrics must be considered, assessing the classification performance from different points of view [69]. The scientific literature presents several dozen performance metrics for classification based on threshold, probabilities or ranks [69]–[75].

The following performance metrics have been implemented in this study: Accuracy, Precision (or Specificity), Recall, Balanced Accuracy & AUC, and F1-score.

Confusion matrix

The classification model predicts the class of each instance, where each sample falls under one of the four cases at the end of the procedure:

- True Positives (TP) are actual positives that have been predicted positive.
- True Negatives (FN) are actual negatives that have been predicted negative.
- False Positives (FP) are actual negatives that have been predicted as positive.
- False Negatives (FN) are actual positives that have been predicted as negative.

Where, in the case of predicting the state of space heating, positives are the "On" labels and negatives are the "Off" labels. A confusion matrix is a table that completely describes the outcome of the classification, showing the resulting number of each of the four different cases (Figure 10).

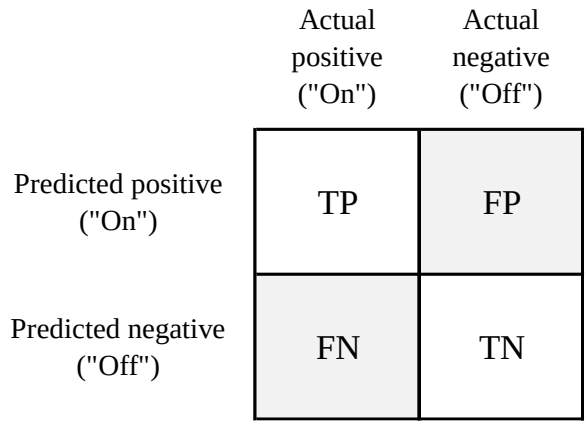


Figure 10. The confusion matrix.

Accuracy

Accuracy is the most simple and common performance metric, which can be easily derived from the confusion matrix and represents the percentage of correctly predicted instances out of all of the instances in the dataset:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

However, according to the empirical evidence, the accuracy metric is only useful in well-balanced datasets as it can be biased to the data imbalance and proportions of incorrect and correct classifications [74].

Recall

Otherwise referred to as sensitivity or the true positive rate, *recall* represents the percentage of positive instances that are correctly predicted:

$$Recall = \frac{TP}{TP + FN}$$

The important feature of the recall metric is that it measures how many positive cases the algorithm picks up, which can be the primary aim of the model in some fields [73]. Recall can be a good measure for imbalanced classification as it focuses only on the positive instances.

Precision

Precision (or Confidence) represents the percentage of predicted positive cases that are correctly predicted:

$$Precision = \frac{TP}{TP + FP}$$

In other words, precision acts as a measure of the accuracy of predicted positives.

Balanced Accuracy & AUC

Balanced Accuracy is based on the sensitivity (true positive rate) and its counterpart specificity (true negative rate), which represents the percentage of negative instances that are correctly predicted, therefore making the balanced accuracy metric especially useful for imbalanced classes. Balanced accuracy is found by taking the arithmetic mean of sensitivity and specificity:

$$Balanced\ Accuracy = \frac{Sensitivity + Specificity}{2}$$

AUC (Area Under the Curve) is a popular ranking type metric for imbalanced datasets which is a quantitative representation of the Receiver Operating Characteristic (ROC) curve. The ROC is a probability curve that depicts the trade-off between the benefits (true positive rate) and costs (false positive rate) of the model. For the binary case, the AUC is equivalent to the Balanced Accuracy [76].

However, Balanced Accuracy and AUC are suboptimal for heavily imbalanced data, as the false positive rate can be reduced due to a large number of true negatives.

F1-score

F1-score is a metric that represents a balanced combination of the model's true positive rate (*Recall*) and the accuracy of the predicted positives (*Precision*). *F1-score* is calculated by taking the harmonic mean of precision and recall:

$$F_1 = 2\frac{Precision \times Recall}{Precision + Recall}$$

F1 is a good metric for imbalanced datasets as it keeps the balance between precision and recall and is sensitive to one of the two inputs having a low value. However, F1 only considers the positive instances. As such, in cases where the imbalanced dataset requires attention to the negatives as well as positives, the Balanced Accuracy is a more appropriate metric.

4 Results and Analysis

This chapter includes 5 sections that will present the results of the methodology described in the previous chapter and their analysis. Section 4.1 will discuss the trends and seasonality in the training dataset. Section 4.2 will present the heating state ground truth labels and discuss their accuracy. Section 4.3 will present the results of classification models and discuss the performance of the classifiers and the importance of the predictors.

4.1 Trends and seasonality

Domestic energy consumption is known to follow seasonal patterns, with common assumptions of increased heating during winter months, as well as daily patterns with morning and evening consumption peaks [17]. The training dataset used in this study was explored to compare against such assumptions.

Figure 11 shows the average monthly external temperatures and gas consumption for 2017. A clear annual trend can be observed with the reversed relationship between external temperatures and gas consumption. The gas consumption is severely reduced from May to September, which is most likely associated with reduced heating demand in the warmer months. Therefore, the extracted features described in section 3.4.1 that reflect the seasonality (*yday*, *week*, *mnth*, *qrtr*) can be considered informative for heating state classification.

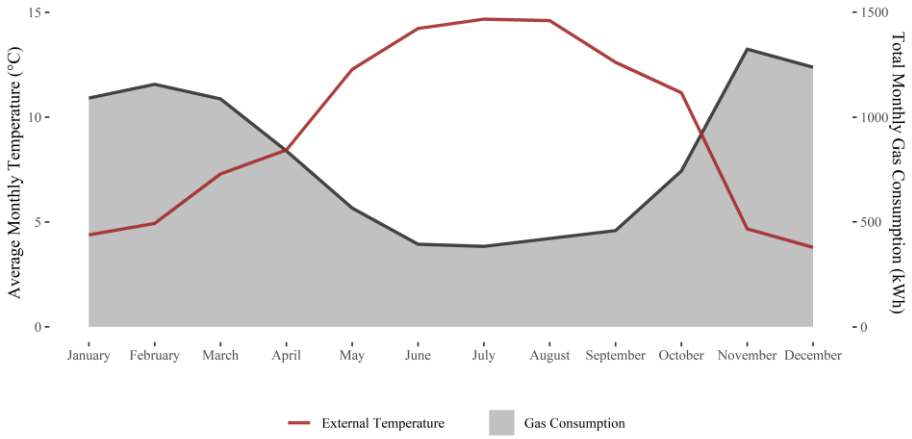


Figure 11. Average monthly external temperatures and total monthly gas consumption averages across all homes in the training dataset in 2017.

Moreover, when comparing the consumption pattern in the first halves of 2017 and 2018, higher consumption values can be observed in 2018, which is likely to be associated with lower external temperatures in that year (Figure 12). Both this and the seasonal variation supports that the external temperature can be an important predictor for heating use.

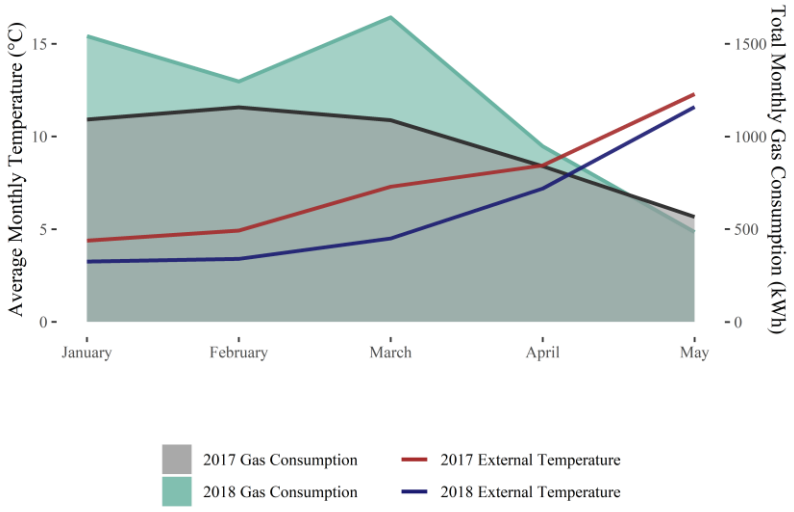


Figure 12. Average monthly temperatures and total monthly consumption averages between all of the homes in the training dataset in the first halves of 2017 and 2018.

Figure 13 shows the average daily temperature and average half-hourly gas consumption on different days of the week in 2017. As expected the external temperatures don' t show any significant correlation. However, the gas consumption is larger during the week, than on weekends, with the largest value on Fridays. This contradicts the common modelling consumption of longer heating periods on weekends or findings of empirical studies that show similar heating periods throughout the week [15], [17], [18].

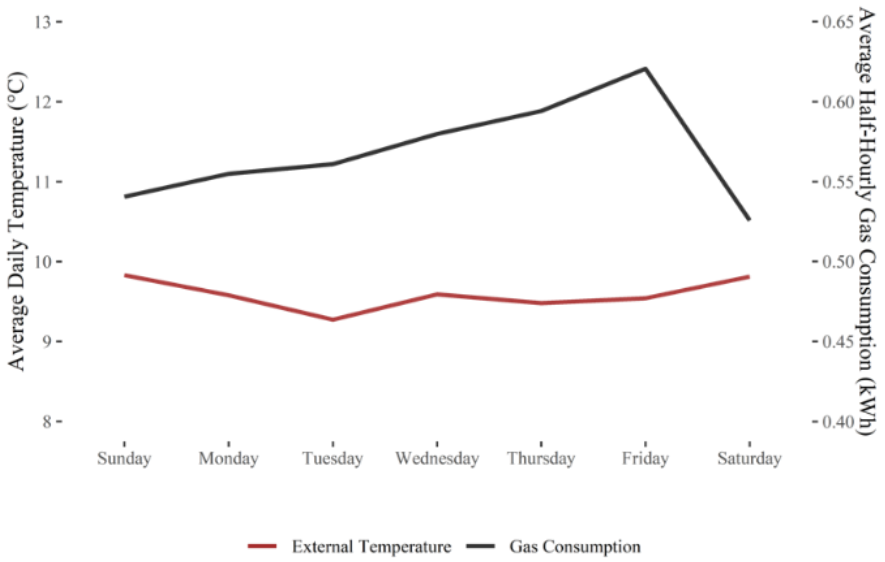


Figure 13. Average daily temperature and average half-hourly gas consumption on different days of the week in 2017 across all homes in the training dataset.

The daily consumption pattern is explored in Figure 14, showing the average half-hourly gas consumption for every hour of the day in 2017. The data shows two clear consumption peaks, one between 6 am and 9 am, and another between around 5 pm and 8 pm. This supports the assumptions and findings from the literature [15], [17], [18].

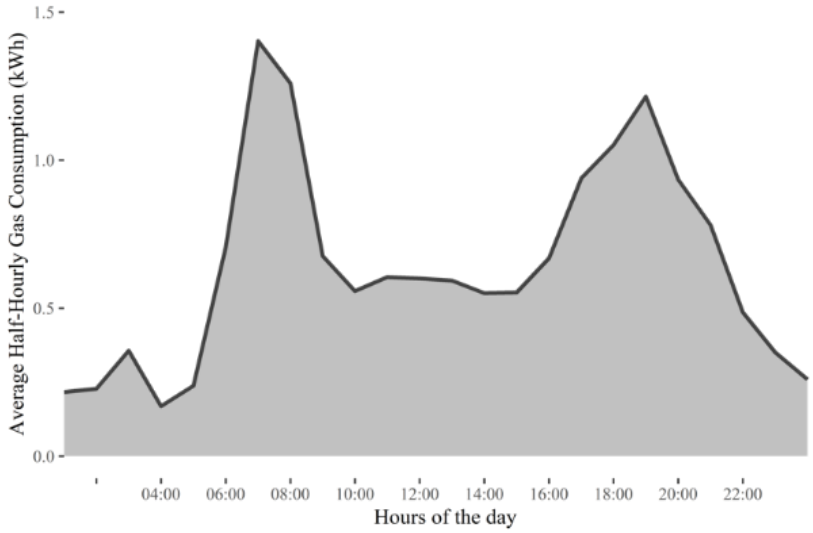


Figure 14. Average half-hourly gas consumption for every hour of the day in 2017 across all homes in the training dataset.

Moreover, large uncertainty between the homes reported in the empirical studies [15], [17], [18] can be observed to an extent in the training sample (Figure 15). Although most homes follow a pattern with 6 am to 9 am and 5 pm to 8 pm peaks, some homes display much higher consumption and more prominent peaks than others, and some show increased consumption throughout the day. It should be noted that the explored data is only a small sample with all homes located in the same area and a larger uncertainty is likely to be expected in a larger sample.

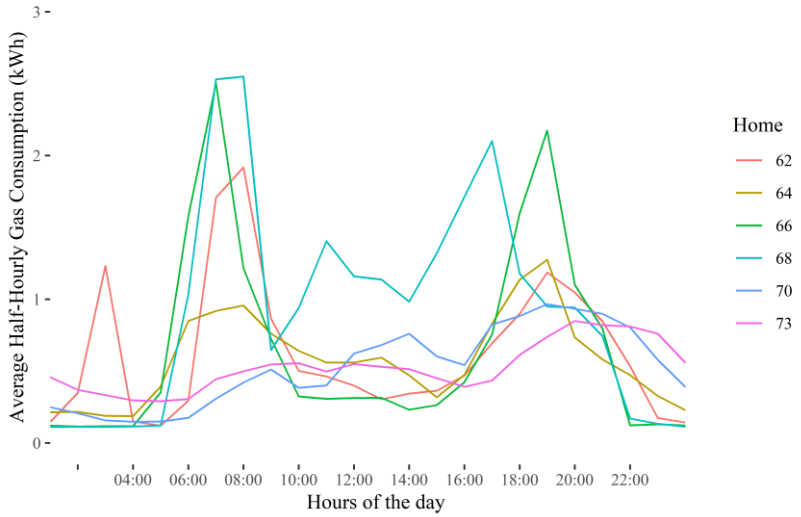


Figure 15. Average half-hourly gas consumption for every hour of the day in 2017 for each home in the training dataset.

4.2 Ground truth labels

To treat the gas consumption disaggregation as a classification problem, three sets of ground truth labels *L1*, *L2* and *L3* were generated by the three methods *M1*, *M2* and *M3* described in section 3.3. *M1* produced labels from living room temperatures, *M2* used average indoor temperatures and *M3* produced labels from heating flow temperatures (Figure 16).

Figure 16. Example day from Home 62 showing how the gas pulse and heating flow temperature the labels generated by three methods; *L1*, *L2* and *L3* are the labels generated by *M1*, *M2* and *M3* respectively.

Upon initial visual inspection, the three sets of labels seem to successfully identify active space heating, avoiding smaller peaks in gas pulse and heating flow temperatures that are likely to be associated with water heating. However, some significant discrepancies are present in the period of active space heating identified. These discrepancies are visible in Figure 16 and are present throughout the data.

Labels *L3* capture the periods of high values in heating flow temperatures, which is to be expected, considering that *L3* is produced using method *M3*,

which takes the heating flow temperatures as an input. However, there can be seen a delay between the onsets of peaks in gas pulse and heating flow temperature (Figure 17), which is also present in the labels *L3*. This delay is likely to be explained by the time required for the water to get heated by gas in the boiler and is usually within 30 minutes in duration.

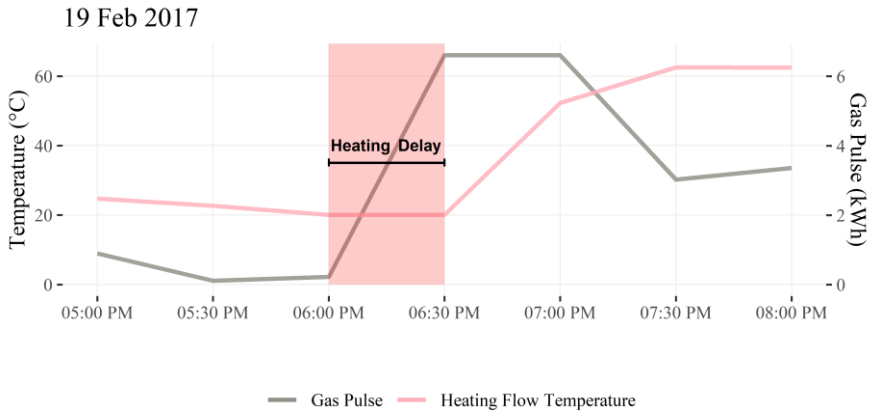


Figure 17. Example from Home 62 showing the heating delay: delay between the onsets of peaks of gas pulse and heating flow temperature associated with the time needed for water to get heated by gas in the boiler.

Labels *L1*, generated from the living room temperatures, often also seem to be delayed from the gas peak onset by 30 minutes. Moreover, the *L1* labels often include only the initial 30 minutes of the heating flow temperature peak, missing the prolonged periods of boiler cycling that are associated with diminished gas consumption, which can be observed in the evening peak in Figure 16.

In the contrast, labels *L2*, generated from the average indoor temperatures, include the full period covering the peaks of gas pulse and heating flow temperature with prolonged boiler cycling periods from onset to offset. This method is likely to be optimal for the majority of the applications, such as understanding energy flexibility or identifying peak consumption periods. Understanding energy flexibility requires accurate identification of offsets of the heating flow temperature peaks, as they indicate when the hot water supply to the radiators has stopped [44]. Understanding peak demand requires accurate identification of both onsets and offsets of gas consumption peaks [77].

The *L1* and *L2* labels were further validated against the *L3* labels using a set of classification performance metrics (Table 3).

Table 3. Performance metric results of L1 and L2 labels validated against L3 labels.

	Home 62		Home 63		Home 64		Home 66		Home 67		Home 68		Home 70		Home 73	
	L1	**L2**	**L1**	**L2**	**L1**	**L2**	**L1**	**L2**	**L1**	**L2**	**L1**	**L2**	**L1**	**L2**	**L1**	**L2**
Accuracy	94%	96%	93%	92%	96%	92%	92%	94%	94%	93%	94%	90%	92%	93%	86%	88%
Precision	68%	74%	73%	70%	41%	24%	61%	69%	80%	73%	85%	61%	39%	43%	54%	72%
Recall	51%	83%	66%	59%	22%	57%	62%	70%	77%	82%	69%	75%	94%	96%	9%	25%
F1	**58%**	**78%**	**69%**	**64%**	**28%**	**34%**	**62%**	**69%**	**78%**	**78%**	**76%**	**67%**	**55%**	**60%**	**16%**	**38%**
Balanced Accuracy & AUC	74%	90%	81%	78%	60%	75%	79%	83%	87%	89%	84%	84%	93%	95%	54%	62%

Both labels show very high accuracies across all homes. However, the labels are heavily skewed, with more occurrences of the "Off" class (Figure 18). Therefore, the high accuracies for this case are most likely misleading and other metrics such as the Balanced Accuracy and F1-score should be considered instead.

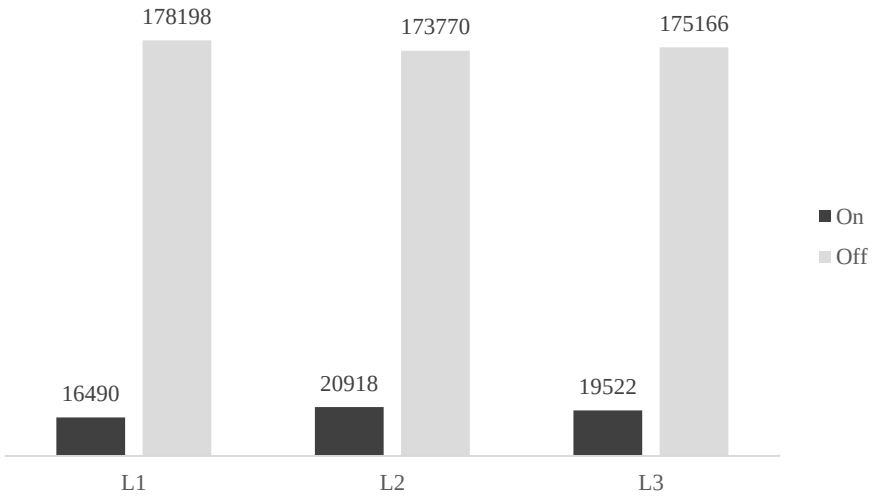


Figure 18. The number of occurrences of "Off" and "On" classes across all generated labels. Class "On" constitutes only 9%, 12% and 11% in *L1*, *L2*, and *L3* respectively.

Moreover, according to *L3*, the proportion of heating operation labels during weekdays and weekends is similar, with heating being "On" 9.6% of the time on weekdays and 10% of the time during the weekend (Figure 19). This contradicts the average half-hourly gas consumption in Figure 13, however, falls in line with the findings in empirical literature [15], [17], [18]. This could be explained by increased gas consumption for uses other than space heating, such as more frequent water heating or cooking during the week.

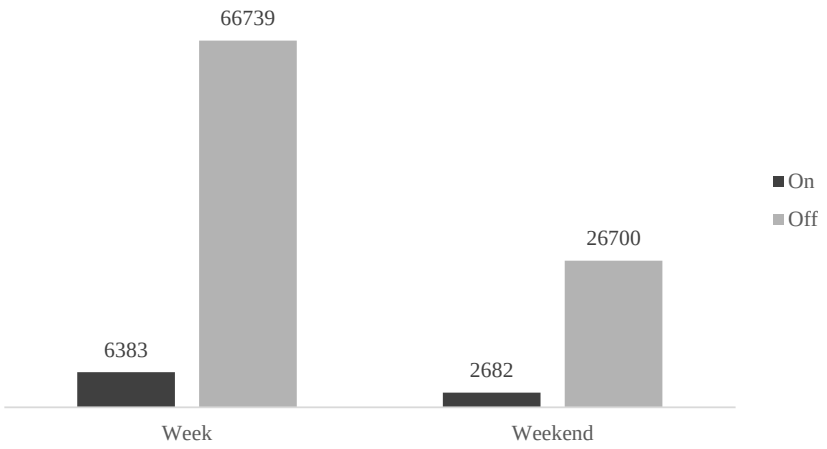


Figure 19. The number of occurrences of "Off" and "On" classes in L3 labels during the weekdays and the weekend.

The Balanced Accuracy and F1-score show predominantly good performance, with exception of homes 64 and 73. The poor performance in homes 64 and 73 could be explained by unusual occupancy patterns in both homes. Home 64 is occupied 6 days a week and home 73 a full week, while the rest of the homes are occupied no more than 4 days a week. A higher score in Balanced Accuracies is likely to be caused by a large number of true negatives, which is associated with heavily imbalanced data. As for this work, prediction of the "On" class is more important, the F1-scores are considered to be more relevant. Comparing the average Balanced Accuracies and F1-scores it is clear that the *L2* labels are significantly more robust than *L1* labels (Table 4).

Table 4. Average performance metric results of L1 and L2 labels validated against L3 labels.

	L1	L2
Accuracy	93%	92%
Precision	63%	61%
Recall	56%	68%
F1	**55%**	**61%**
Balanced Accuracy & AUC	77%	82%

To reduce the imbalance in the dataset, the data was filtered to only include the heating months (October to April), with more frequent heating events, and consequently, a slightly more prevalent "On" class (Figure 20).

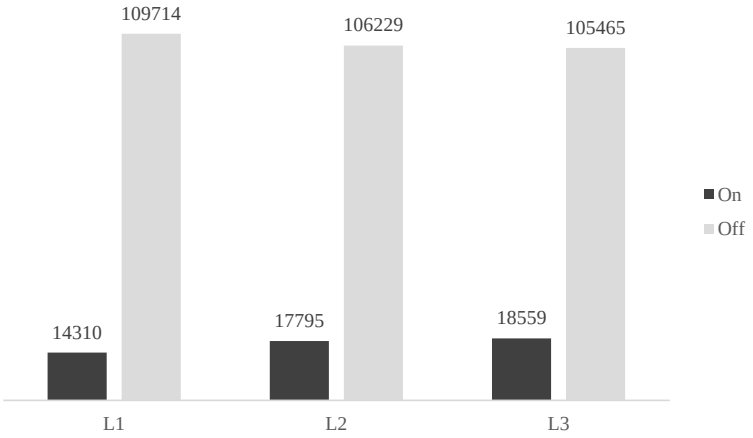


Figure 20. The number of occurrences of "Off" and "On" across all generated labels using filtered data. Class "On" constitutes 13%, 17% and 18% in *L1*, *L2*, *L3* respectively.

The *L1* and *L2* labels have been validated against *L3* labels once more but using the data from heating months (Table 5). Despite the losses in accuracy, the F1-scores show significant improvements, which can mainly be explained by the increases in Precision. This means that the algorithms are more robust during the winter months and the positive predictions are likely to be more correct more often. One possible explanation is the tendency to misidentify heat gains during warmer months as heating.

Table 5. (a) The performance metric results of filtered data (*L1* and *L2* labels validated against L3 labels); (b) difference in the performance of the labels generated from the filtered data and labels generated from the full data.

a

	L1	L2
Accuracy	90%	90%
Precision	67%	67%
Recall	57%	69%
F1	**58%**	**65%**
Balanced Accuracy & AUC	77%	82%

b

L1	L2
-3%	-2%
4%	6%
1%	1%
3%	**4%**
0%	0%

Overall, the metrics show that the *L2* labels, generated from the average indoor temperatures have a higher correlation with the *L3* labels generated from the heating flow temperatures. Despite the living room being conventionally the most heated in the dwelling in terms of frequency and temperature [78], there might often be periods of heating events happening only in one of the other rooms. During these periods the *L1* labels, generated from the living room temperatures, would be classed as "Off". Using average temperatures allows for capturing changes in all of the rooms, making it a more robust source of labels, albeit less sensitive. Therefore, the method *M2* is considered more robust and the *L2* labels produced by this method are used as ground truth in further modelling.

4.3 Classification

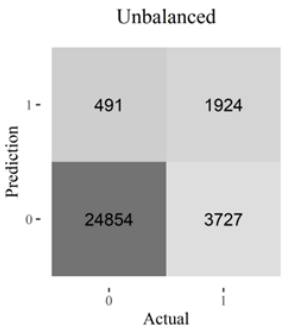
Four classification algorithms have been tested on the dataset using the labels generated in the previous section as ground truth: Logistic Regression (LR), Decision Trees (DT), Random Forest (RF) and XGBoost (XGB). To predict the state of the heating system the models employed 21 predictors described in section 3.4.1: *gas*, *t*, *hr*, *yday*, *mday*, *wday*, *wend*, *week*, *mnth*, *qrtr*, *l1*, *l2*, *l3*, *l4*, *l5*, *rmean*, *rmed*, *rstd*, *rslope*, *rskew*, *rkurt*.

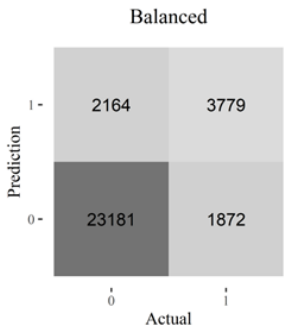
The initial dataset was split into train, validation and test datasets, where train and validation contain the data collected from the same homes but in different years, while the test dataset contains data from different homes. Results presented for each model show performance of classification on validation and test datasets of models trained on both unbalanced training data and training data balanced via oversampling.

4.3.1 Logistic regression

Figure 21 shows the performance of the Logistic Regression model. The model trained on the unbalanced dataset achieved high performance on validation data, but low performance on test data. This is likely due to overfitting and bias towards the negative labels. The discrepancy between the Balanced Accuracy and F1-score is also likely to be explained by the imbalance in the dataset, with a larger number of negative labels raising the Balanced Accuracy.

The model trained on the balanced dataset achieves lower performance on validation data. Although there is also a decrease in precision of the balanced model on test data, the significant increase in recall results in a much higher F1-score improving the overall performance. This shows that balancing the data helped to solve the overfit and bias. The resulting F1-score on test data for the model trained using a balanced dataset was 65%.





	Validation	Test
Accuracy	92%	86%
Precision	74%	80%
Recall	67%	34%
F1	**71%**	**48%**
Balanced Accuracy & AUC	82%	66%

	Validation	Test	
	86%	87%	**Accuracy**
	50%	64%	**Precision**
	87%	67%	**Recall**
	64%	**65%**	**F1**
	86%	79%	**Balanced Accuracy & AUC**

Figure 21. Confusion matrix and performance metrics of Logistic Regression on validation and test datasets, trained on unbalanced and balanced data.

ROC curve analysis was conducted to estimate the variable importance, finding the AUC for a series of cut-offs applied to the predictor data (Figure 22). Gas consumption and its lagged variables were found to be of significance, as well as the temperature and a few date-time features representing the day of the week, the month of the year and the hour of the day.

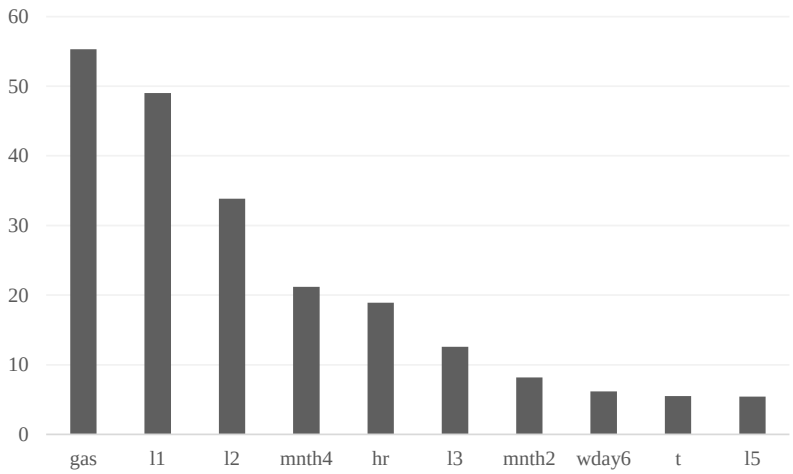
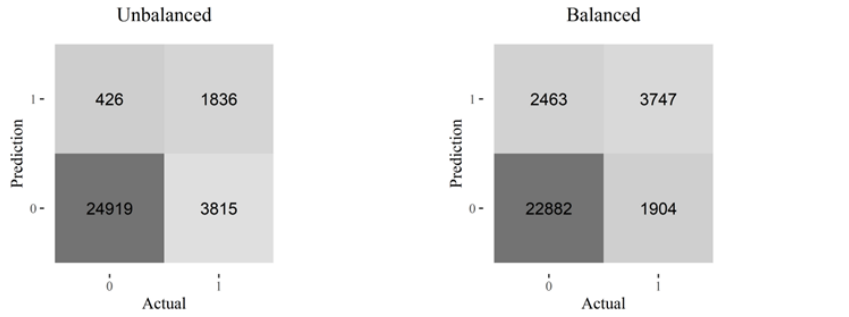


Figure 22. Variable importance in Logistic Regression model.

4.3.2 Decision Tree

Similarly, the overfitting in the Decision Tree model was also fixed by balancing the training dataset (Figure 23). The scores achieved by the model are very similar to the Logistic Regression model score, although the decision tree model is performing slightly worse in the majority of metrics, with the model trained on balanced data achieving an F1-score of 63% in test data.



	Validation	Test
Accuracy	93%	86%
Precision	76%	81%
Recall	69%	32%
F1	**72%**	**46%**
Balanced Accuracy & AUC	83%	65%

	Validation	Test
Accuracy	85%	86%
Precision	48%	60%
Recall	88%	66%
F1	**62%**	**63%**
Balanced Accuracy & AUC	86%	78%

Figure 23. Confusion matrix and performance metrics of Decision Tree model on validation and test datasets, trained on unbalanced and balanced data.

The Decision Tree was built using default parameters in the *rpart* package [79]:

- The maximum depth of the tree was set to 30.
- The minimum number of examples in a node to perform a split was set to 20.
- The minimum number of examples in the terminal node was set to 7.

The resulting Tree is displayed in Figure 24. Lagged variables and rolling statistics extracted from gas consumption data are used for splitting the tree, with a rolling slope at the first split. A large portion of the data is also given a negative label only after the second split, which used the second lagged variable as a threshold. The tree could be interpreted as simply classifying larger gas consumption values as used for space heating.

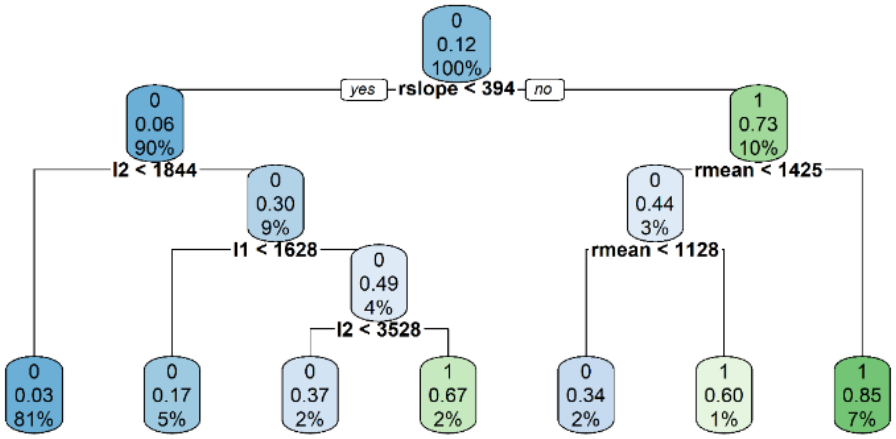


Figure 24. The decision tree is built using default parameters.

To improve the performance of the model, the Decision Tree hyperparameters have been tested iteratively using F1-score as a metric. Following 3000 iterations the resulting parameters were:

- The maximum depth of the tree was set to 5.
- The minimum number of examples in a node to perform a split was set to 3.
- The minimum number of examples in the terminal node was set to 7.

The resulting tree is displayed in Figure 25. Although the majority of splits are still made on rolling statistics and lagged variables of gas consumption, some of the splits are also made using the feature representing the hour of the day. For smaller values of gas consumption heating before 1 pm and for larger gas consumption values heating before 7 am is considered less likely.

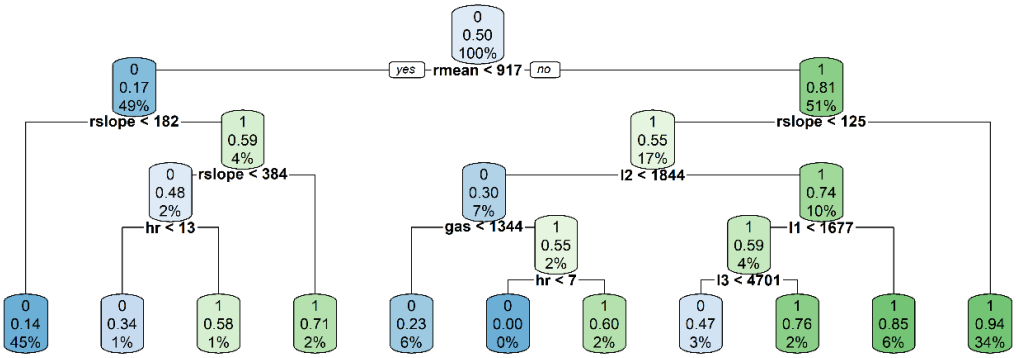
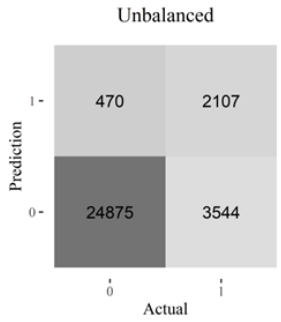
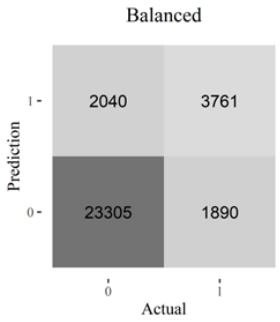


Figure 25. A decision tree built using tuned parameters.

Tuning the parameters has improved the model performance, achieving better scores in models trained both on unbalanced and balanced data (Figure 26). The tuned model also outperformed the Logistic Regression model in the majority of metrics, achieving an F1-score of 66% on test data in the model trained on the balanced dataset.





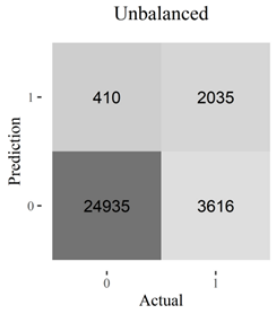
	Validation	Test
Accuracy	93%	87%
Precision	75%	82%
Recall	71%	37%
F1	**73%**	**51%**
Balanced Accuracy & AUC	83%	68%

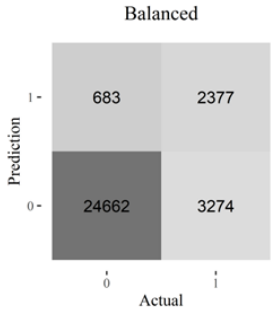
Validation	Test	
87%	87%	**Accuracy**
52%	65%	**Precision**
87%	67%	**Recall**
65%	**66%**	**F1**
87%	79%	**Balanced Accuracy & AUC**

Figure 26. Confusion matrix and performance metrics of tuned decision tree model on validation and test datasets, trained on unbalanced and balanced data.

4.3.3 Random Forest

Figure 27 shows the performance of the Random Forest model. The model trained on unbalanced data performs better than the respective Decision Tree and Logistic Regression models. However, training the model on balanced data results only in a slight increase in model performance, achieving much smaller scores than previous models. The F1-score of the test data classification by the Random Forest trained on balanced data is only 55%.





	Validation	Test
Accuracy	93%	87%
Precision	77%	83%
Recall	74%	36%
F1	**76%**	**50%**
Balanced Accuracy & AUC	85%	67%

	Validation	Test
Accuracy	92%	87%
Precision	69%	78%
Recall	78%	42%
F1	**73%**	**55%**
Balanced Accuracy & AUC	86%	70%

Figure 27. Confusion matrix and performance metrics of random forest model on validation and test datasets, trained on unbalanced and balanced data.

The model parameters are set as:

- The number of trees to grow is set to be 500 [80].
- The number of variables randomly sampled as candidates at each split is taken as the square root of the total number of predictors [80].

10-fold cross-validation with 3 repeats is also performed to limit and reduce overfitting [81]. The parameters were not further tuned due to the high computational cost.

ROC curve analysis shows that similarly to Decision Tree models, the rolling statistics and lagged variables of gas consumption are of significance to the model (Figure 28). Other important variables include date-time features such as the hour of the day and the day of the year.

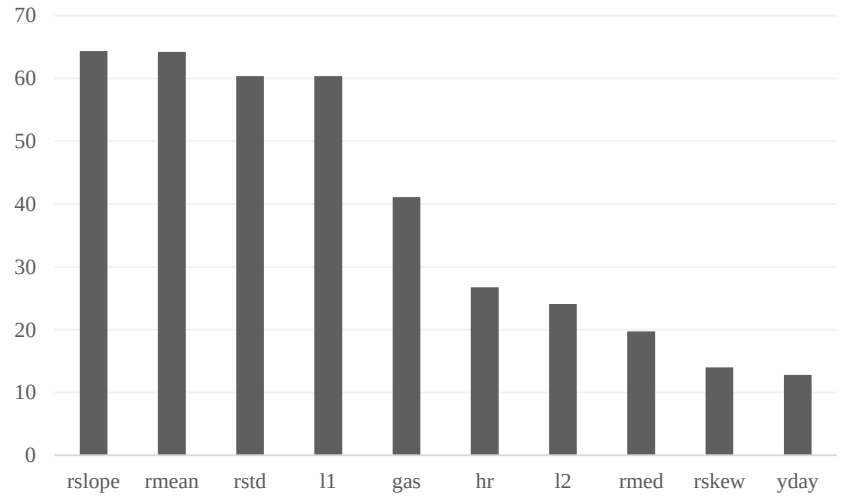
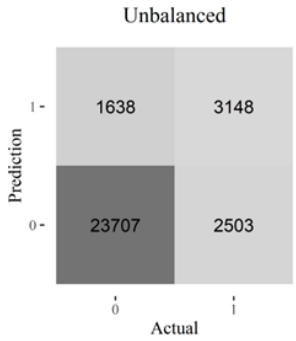
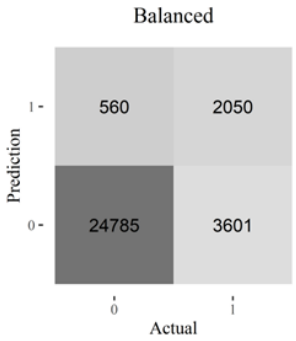


Figure 28. Variable importance in Random Forest model.

4.3.4 XGBoost

Figure 29 shows the performance of the XGBoost model. Similarly to the Random Forest model, XGBoost performs well with the unbalanced training dataset, however, does not improve as much when oversampling the minority class for training. Despite that, the model trained on the balanced data performs better than its Random Forest counterpart, with an F1-score of 60%.





	Validation	Test
Accuracy	97%	87%
Precision	95%	79%
Recall	81%	36%
F1	**88%**	**50%**
Balanced Accuracy & AUC	90%	67%

	Validation	Test
Accuracy	96%	87%
Precision	95%	66%
Recall	98%	56%
F1	**96%**	**60%**
Balanced Accuracy & AUC	96%	75%

Figure 29. Confusion matrix and performance metrics of random forest model on validation and test datasets, trained on unbalanced and balanced data.

The model is trained using the default parameters [82]:

- The maximum depth for a tree was selected to be 6.
- The minimum sum of weights of all observations required in a child as well as the fraction of the observation to be randomly sampled for each tree was selected to be 1.

The parameters were not further tuned due to the computational costs.

The feature importance in the model was calculated by the information gain reduction of each node after splitting using the variable (Figure 30). Similarly to previous findings, the rolling means and lagged variables of gas consumption are of significance as well as the feature representing the hour of the day.

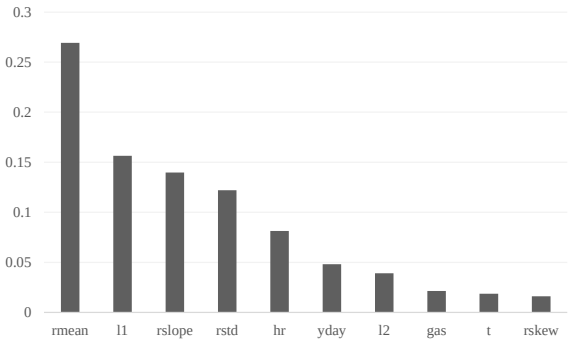


Figure 30. Variable importance in XGBoost model.

5 Conclusion and Future Work

Identifying heating periods in gas consumption data could be crucial in facilitating the transition to renewable energy sources in the domestic sector. This study proposes a methodology that would allow the identification of heating state in homes using gas meter data and external temperatures that are easier and cheaper to collect by training a classification model with ground truth labels generated from a smaller sample of collected indoor temperature data.

The sample used in this study shows trends and seasonality in gas consumption. Despite a large variation between homes, there is a general daily pattern with two consumption peaks, one from 6 am to 9 am and another from 5 pm to 8 pm. The data also shows increased gas consumption during the weekdays, despite the same amount of heating hours identified throughout the week. Increased gas consumption by different end uses could explain this discrepancy.

Ground Truth Labels

Heating state labels generated from living room temperatures and average indoor temperatures are validated by heating flow temperatures and are found to show correct predictions. However, the labels generated from average indoor temperatures are found to be most accurate for three reasons:

- The labels show high performance when validated against labels generated from heating flow temperatures. This signifies that the labels are less susceptible to misidentifying variation in temperatures caused by secondary factors, such as heat gains or losses from external temperatures, window opening, secondary heating or cooking.
- Using average temperatures allows us to identify temperature changes throughout the house, which would include heating operations that happen only outside the living room, and that is not captured by the living room temperatures.

- The labels are observed to cover the full period that is associated with heating, including the prolonged periods of boiler cycling.

Despite the literature suggesting employing the radiator temperatures as the most robust source of inferred heating labels, such data is harder and more expensive to find, while indoor temperatures could be available from smart thermostats. However, it should be noted that this study achieves robust results using average indoor temperatures, while a bigger sample of temperatures collected from thermostats would include larger uncertainty due to the thermostat being located in different rooms.

As would be expected, the heating labels were found to be heavily imbalanced, with the majority in the "Off" class. This renders the accuracy metric meaningless and study considers the F1-score as the primary metric as it focuses on positive ("On") labels, which are important for the purposes of this study.

The labels were also found to perform better during the winter months, which supports the previous findings in the literature.

Classification

The labels generated from average indoor temperatures were further used as ground truth in classification, predicting heating operation using only gas meter data and external temperatures, as well as temporal features extracted from the gas meter data.

All four tested algorithms perform poorly due to the training dataset being highly imbalanced, creating overfit and bias towards negative ("Off") labels. Oversampling the minority ("On") class improved the performance of the models, solving the bias and overfit in some of them. The best performing algorithms were Logistic Regression, with an F1-score on test data equal to 65% and Decision Tree with tuned parameters, with an F1-score on test data equal to 66%.

In contrast, although more complex and theoretically superior, XGBoost and Random Forest did not improve as much from balancing the training dataset and achieved poor performance scores. The F1-score achieved by Random Forest on test data was 55%, while XGBoost achieved 60%. That said, both Random Forest and XGBoost were trained using the default parameters and could improve in performance if the parameters were tuned.

The variable significance was calculated in the models, finding variables of rolling statistics extracted from the gas consumption by far the most significant, followed by the gas consumption and its lagged variables. A date-time feature that represented the time of the day was also found to be important. The other date-time variables were not as significant.

Future work

Although the performance metric scores achieved by best-performing classifiers are not very high, there is room for improvement.

- Parameter tuning in XGBoost and Random Forest is likely to improve the performance.
- Different classification algorithms, including support vector machines and neural networks, could be a better fit for this problem.
- Methods of balancing the training dataset, other than oversampling, could be explored.
- The study uses only a small subset of homes. Employing more data is more than likely to have a positive impact on model performance.
- Using dwelling and resident characteristics as predictors could address a lot of variability in consumption and heating patterns between different homes.
- Using direct measurements from heating controls could help improve the ground truth label generation accuracy.

Bibliography

[1] CCC, "The Sixth Carbon Budget Buildings," 2020, Accessed: Jun. 08, 2022. [Online]. Available: www.theccc.org.uk

[2] HMG, "The Climate Change Act 2008 (2050 Target Amendment) Order 2019," 2019. https://www.legislation.gov.uk/ukdsi/2019/9780111187654 (accessed Jun. 08, 2022).

[3] BPIE, "Europe's buildings under the microscope. A country-by-country review of the energy performance of buildings.," 2011.

[4] N. Kerr and M. Winskel, "A review of heat decarbonisation policies in Europe Executive summary," 2021, doi: 10.7488/era/794.

[5] UKGBC, "Retrofit Policy Playbook," 2021.

[6] J. Allison, A. Cowie, S. Galloway, J. Hand, N. J. Kelly, and B. Stephen, "Simulation, implementation and monitoring of heat pump load shifting using a predictive controller," *Energy Convers Manag*, vol. 150, pp. 890–903, Oct. 2017, doi: 10.1016/J.ENCONMAN.2017.04.093.

[7] N. Eyre and P. Baruah, "Uncertainties in future energy demand in UK residential heating," *Energy Policy*, vol. 87, pp. 641–653, Dec. 2015, doi: 10.1016/J.ENPOL.2014.12.030.

[8] I. A. G. Wilson, A. J. R. Rennie, Y. Ding, P. C. Eames, P. J. Hall, and N. J. Kelly, "Historical daily gas and electrical energy flows through Great Britain's transmission networks and the decarbonisation of domestic heat," *Energy Policy*, vol. 61, pp. 301–305, Oct. 2013, doi: 10.1016/J.ENPOL.2013.05.110.

[9] S. Ø. Jensen *et al.*, "IEA EBC Annex 67 Energy Flexible Buildings," *Energy Build*, vol. 155, pp. 25–34, Nov. 2017, doi: 10.1016/J.ENBUILD.2017.08.044.

[10] CCC, "Next Steps for UK Heat Policy," 2016. Accessed: Jun. 08, 2022. [Online]. Available: https://www.theccc.org.uk/publication/next-steps-for-uk-heat-policy/

[11] IEA, "Heat Pumps in Smart Grids Technology Collabortation Programme on Heat Pumping Technologies," 2017. https://heatpumpingtechnologies.org/publications/heat-pumps-in-smart-grids-final-report/ (accessed Jun. 10, 2022).

[12] R. Zafar, A. Mahmood, S. Razzaq, W. Ali, U. Naeem, and K. Shehzad, "Prosumer based energy management and sharing in smart grid," *Renewable and Sustainable Energy Reviews*, vol. 82, pp. 1675–1684, Feb. 2018, doi: 10.1016/J.RSER.2017.07.018.

[13] A. Gillich, M. Saber, and E. Mohareb, "Limits and uncertainty for energy efficiency in the UK housing stock," 2019, doi: 10.1016/j.enpol.2019.110889.

[14] Y. Chen, P. Xu, J. Gu, F. Schmidt, and W. Li, "Measures to improve energy demand flexibility in buildings for demand response (DR): A review," *Energy Build*, vol. 177, pp. 125–139, Oct. 2018, doi: 10.1016/J.ENBUILD.2018.08.003.

[15] G. M. Huebner, M. McMichael, D. Shipworth, M. Shipworth, M. Durand-Daubin, and A. Summerfield, "Heating patterns in English homes: Comparing results from a national survey against common model assumptions," *Build Environ*, vol. 70, pp. 298–305, Dec. 2013, doi: 10.1016/J.BUILDENV.2013.08.028.

[16] M. Kavgic, A. Mavrogianni, D. Mumovic, A. Summerfield, Z. Stevanovic, and M. Djurovic-Petrovic, "A review of bottom-up building stock models for energy consumption in the residential sector," *Build Environ*, vol. 45, no. 7, pp. 1683–1697, Jul. 2010, doi: 10.1016/J.BUILDENV.2010.01.021.

[17] G. M. Huebner, M. McMichael, D. Shipworth, M. Shipworth, M. Durand-Daubin, and A. Summerfield, "The reality of English living rooms – A comparison of internal temperatures against common model assumptions," *Energy Build*, vol. 66, pp. 688–696, Nov. 2013, doi: 10.1016/J.ENBUILD.2013.07.025.

[18] M. Shipworth, S. K. Firth, M. I. Gentry, A. J. Wright, D. T. Shipworth, and K. J. Lomas, "Central heating thermostat settings and timing: building demographics," *https://doi.org/10.1080/09613210903263007*, vol. 38, no. 1, pp. 50–69, Jan. 2009, doi: 10.1080/09613210903263007.

[19] S. Kelly, "Do homes that are more energy efficient consume less energy?: A structural equation model of the English residential sector," *Energy*, vol. 36, no. 9, pp. 5610–5620, Sep. 2011, doi: 10.1016/J.ENERGY.2011.07.009.

[20] P. de Wilde, "The gap between predicted and measured energy performance of buildings: A framework for investigation," *Autom Constr*, vol. 41, pp. 40–49, May 2014, doi: 10.1016/J.AUTCON.2014.02.009.

[21] A. Alzaatreh, L. Mahdjoubi, B. Gething, and F. Sierra, "Disaggregating high-resolution gas metering data using pattern recognition," *Energy Build*, vol. 176, pp. 17–32, Oct. 2018, doi: 10.1016/J.ENBUILD.2018.07.011.

[22] E. Mangematin, G. Pandraud, and D. Roux, "Quick measurements of energy efficiency of buildings," *C R Phys*, vol. 13, no. 4, pp. 383–390, May 2012, doi: 10.1016/J.CRHY.2012.04.001.

[23] S. Hammarsten, "A critical appraisal of energy-signature models," *Appl Energy*, vol. 26, no. 2, pp. 97–110, Jan. 1987, doi: 10.1016/0306-2619(87)90012-2.

[24] A. Afram and F. Janabi-Sharifi, "Theory and applications of HVAC control systems – A review of model predictive control (MPC)," *Build Environ*, vol. 72, pp. 343–355, Feb. 2014, doi: 10.1016/J.BUILDENV.2013.11.016.

[25] M. Dahl Knudsen and S. Petersen, "Demand response potential of model predictive control of space heating based on price and carbon dioxide intensity signals," *Energy Build*, vol. 125, pp. 196–204, Aug. 2016, doi: 10.1016/J.ENBUILD.2016.04.053.

[26] M. Wytock and J. Zico Kolter, "Contextually Supervised Source Separation with Application to Energy Disaggregation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 28, no. 1, Jun. 2014, doi: 10.1609/AAAI.V28I1.8769.

[27] A. Zoha, A. Gluhak, M. A. Imran, and S. Rajasegarar, "Non-Intrusive Load Monitoring Approaches for Disaggregated Energy Sensing: A Survey," *Sensors 2012, Vol. 12, Pages 16838-16866*, vol. 12, no. 12, pp. 16838–16866, Dec. 2012, doi: 10.3390/S121216838.

[28] M. Pullinger *et al.*, "The IDEAL household energy dataset, electricity, gas, contextual sensor data and survey data for 255 UK homes," *Scientific Data 2021 8:1*, vol. 8, no. 1, pp. 1–18, May 2021, doi: 10.1038/s41597-021-00921-y.

[29] M. Zeifman and K. Roth, "Nonintrusive appliance load monitoring: Review and outlook," *IEEE Transactions on Consumer Electronics*, vol. 57, no. 1, pp. 76–84, Feb. 2011, doi: 10.1109/TCE.2011.5735484.

[30] G. W. Hart, "Nonintrusive Appliance Load Monitoring," *Proceedings of the IEEE*, vol. 80, no. 12, pp. 1870–1891, 1992, doi: 10.1109/5.192069.

[31] S. R. Shaw, S. B. Leeb, L. K. Norford, and R. W. Cox, "Nonintrusive load monitoring and diagnostics in power systems," *IEEE Trans Instrum Meas*, vol. 57, no. 7, pp. 1445–1454, 2008, doi: 10.1109/TIM.2008.917179.

[32] S. R. Shaw and C. R. Laughman, "A Kalman-filter spectral envelope preprocessor," *IEEE Trans Instrum Meas*, vol. 56, no. 5, pp. 2010–2017, Oct. 2007, doi: 10.1109/TIM.2007.904475.

[33] Leeb Michael S, LeVan James L Kirtley, Steven B, and J. P. Joseph Sweeney, "Development and Validation of a Transient Event Detector," 1993.

[34] J. Liang, S. K. K. Ng, G. Kendall, and J. W. M. Cheng, "Load signature studypart I: Basic concept, structure, and methodology," *IEEE Transactions on Power Delivery*, vol. 25, no. 2, pp. 551–560, Apr. 2010, doi: 10.1109/TPWRD.2009.2033799.

[35] M. Baranski and J. Voss, "Non-intrusive appliance load monitoring based on an optical sensor," *2003 IEEE Bologna PowerTech - Conference Proceedings*, vol. 4, pp. 267–274, 2003, doi: 10.1109/PTC.2003.1304732.

[36] M. Baranski and J. Voss, "Genetic algorithm for pattern detection in NIALM systems," *Conf Proc IEEE Int Conf Syst Man Cybern*, vol. 4, pp. 3462–3468, 2004, doi: 10.1109/ICSMC.2004.1400878.

[37] L. Farinaccio and R. Zmeureanu, "Using a pattern recognition approach to disaggregate the total electricity consumption in a house into the major end-uses," *Energy Build*, vol. 30, no. 3, pp. 245–259, Aug. 1999, doi: 10.1016/S0378-7788(99)00007-9.

[38] K. A. Nguyen, R. A. Stewart, and H. Zhang, "An intelligent pattern recognition model to automate the categorisation of residential water end-use events," *Environmental Modelling & Software*, vol. 47, pp. 108–127, Sep. 2013, doi: 10.1016/J.ENVSOFT.2013.05.002.

[39] S. R. Vitullo, "DISAGGREGATING TIME SERIES DATA FOR ENERGY CONSUMPTION BY AGGREGATE AND INDIVIDUAL CUSTOMER," 2011.

[40] P. Bacher, P. A. de Saint-Aubain, L. E. Christiansen, and H. Madsen, "Non-parametric method for separating domestic hot water heating spikes and space heating," *Energy Build*, vol. 130, pp. 107–112, Oct. 2016, doi: 10.1016/J.ENBUILD.2016.08.037.

[41] T. Kane, S. K. Firth, T. M. Hassan, and V. Dimitriou, "Heating behaviour in English homes: An assessment of indirect calculation methods," *Energy Build*, vol. 148, pp. 89–105, Aug. 2017, doi: 10.1016/J.ENBUILD.2017.04.059.

[42] T. Kane, "Indoor temperatures in UK dwellings : investigating heating practices using field survey data," 2013, Accessed: Aug. 30, 2022. [Online]. Available: https://hdl.handle.net/2134/12563

[43] T. Kane, S. K. Firth, and K. J. Lomas, "How are UK homes heated? A city-wide, socio-technical survey and implications for energy modelling," *Energy Build*, vol. 86, pp. 817–832, Jan. 2015, doi: 10.1016/J.ENBUILD.2014.10.011.

[44] J. Crawley, D. Manouseli, P. Mallaburn, and C. Elwell, "An Empirical Energy Demand Flexibility Metric for Residential Properties," *Energies 2022, Vol. 15, Page 5304*, vol. 15, no. 14, p. 5304, Jul. 2022, doi: 10.3390/EN15145304.

[45] N. Berliner, M. Pullinger, and N. Goddard, "Inferring room-level use of domestic space heating from room temperature and humidity measurements using a deep, dilated convolutional network," *MethodsX*, vol. 8, p. 101367, Jan. 2021, doi: 10.1016/J.MEX.2021.101367.

[46] G. K. F. Tso and K. K. W. Yau, "A study of domestic energy usage patterns in Hong Kong," *Energy*, vol. 28, no. 15, pp. 1671–1682, Dec. 2003, doi: 10.1016/S0360-5442(03)00153-1.

[47] ESESCR, "The Edinburgh and South East Scotland City Region Deal," 2022. https://esescityregiondeal.org.uk/ (accessed Aug. 20, 2022).

[48] U. Ali, M. H. Shamsi, C. Hoare, E. Mangina, and J. O' Donnell, "A data-driven approach for multi-scale building archetypes development," *Energy Build*, vol. 202, p. 109364, Nov. 2019, doi: 10.1016/J.ENBUILD.2019.109364.

[49] Y. G. Yohanis, J. D. Mondol, A. Wright, and B. Norton, "Real-life energy use in the UK: How occupancy and dwelling characteristics affect domestic electricity use," *Energy Build*, vol. 40, no. 6, pp. 1053–1059, Jan. 2008, doi: 10.1016/J.ENBUILD.2007.09.001.

[50] BEIS, "Q1 2022 Smart Meters Statistics Report," 2022.

[51] H. Zhou, K. M. Yu, M. G. Lee, and C. C. Han, "The Application of Last Observation Carried Forward Method for Missing Data Estimation in the Context of Industrial Wireless Sensor Networks," *Proceedings of the 2018 IEEE 7th Asia-Pacific Conference on Antennas and Propagation, APCAP 2018*, pp. 130–131, Nov. 2018, doi: 10.1109/APCAP.2018.8538147.

[52] A. Reguis, B. Vand, and J. Currie, "Challenges for the Transition to Low-Temperature Heat in the UK: A Review," *Energies 2021, Vol. 14, Page 7181*, vol. 14, no. 21, p. 7181, Nov. 2021, doi: 10.3390/EN14217181.

[53] E. M. Knorr and R. T. Ng, "A Unified Notion of Outliers: Properties and Computation," 1997, Accessed: Aug. 30, 2022. [Online]. Available: www.aaai.org

[54] G. Petneházi, "Recurrent Neural Networks for Time Series Forecasting," *Cornell University*, pp. 1–22, Jan. 2019, doi: 10.48550/arxiv.1901.00069.

[55] C. L. Liu, W. H. Hsaio, and Y. C. Tu, "Time Series Classification with Multivariate Convolutional Neural Network," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 6, pp. 4788–4797, Jun. 2019, doi: 10.1109/TIE.2018.2864702.

[56] C. Che, H. Wang, X. Ni, and R. Lin, "Hybrid multimodal fusion with deep learning for rolling bearing fault diagnosis," *Measurement*, vol. 173, p. 108655, Mar. 2021, doi: 10.1016/J.MEASUREMENT.2020.108655.

[57] H. Deng, G. Runger, E. Tuv, and M. Vladimir, "A Time Series Forest for Classification and Feature Extraction," *Inf Sci (N Y)*, vol. 239, pp. 142–153, Feb. 2013, doi: 10.48550/arxiv.1302.2277.

[58] F. Shaikh and Q. Ji, "Forecasting natural gas demand in China: Logistic modelling analysis," *International Journal of Electrical Power & Energy Systems*, vol. 77, pp. 25–32, May 2016, doi: 10.1016/J.IJEPES.2015.11.013.

[59] A. Lawi, S. la Wungo, and S. Manjang, "Identifying irregularity electricity usage of customer behaviors using logistic regression and linear discriminant analysis," *Proceeding - 2017 3rd International Conference on Science in Information Technology: Theory and Application of IT for Education, Industry and Society in Big Data Era, ICSITech 2017*, vol. 2018-January, pp. 552–557, Jul. 2017, doi: 10.1109/ICSITECH.2017.8257174.

[60] J. R. Quinlan, "Induction of decision trees," *Machine Learning 1986 1:1*, vol. 1, no. 1, pp. 81–106, Mar. 1986, doi: 10.1007/BF00116251.

[61] Z. Yu, F. Haghighat, B. C. M. Fung, and H. Yoshino, "A decision tree method for building energy demand modeling," *Energy Build*, vol. 42, no. 10, pp. 1637–1646, Oct. 2010, doi: 10.1016/J.ENBUILD.2010.04.006.

[62] G. K. F. Tso and K. K. W. Yau, "Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks," *Energy*, vol. 32, no. 9, pp. 1761–1768, Sep. 2007, doi: 10.1016/J.ENERGY.2006.11.010.

[63] V. Svetnik, A. Liaw, C. Tong, J. Christopher Culberson, R. P. Sheridan, and B. P. Feuston, "Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling," *J Chem Inf Comput Sci*, vol. 43, no. 6, pp. 1947–1958, Nov. 2003, doi: 10.1021/CI034160G/SUPPL_FILE/CI034160GSI20031008_041202.ZIP.

[64] Z. Wang, Y. Wang, R. Zeng, R. S. Srinivasan, and S. Ahrentzen, "Random Forest based hourly building energy prediction," *Energy Build*, vol. 171, pp. 11–25, Jul. 2018, doi: 10.1016/J.ENBUILD.2018.04.008.

[65] A. Zakariazadeh, "Smart meter data classification using optimized random forest algorithm," *ISA Trans*, vol. 126, pp. 361–369, Jul. 2022, doi: 10.1016/J.ISATRA.2021.07.051.

[66] W. Zhang, H. Quan, and D. Srinivasan, "Parallel and reliable probabilistic load forecasting via quantile regression forest and quantile determination," *Energy*, vol. 160, pp. 810–819, Oct. 2018, doi: 10.1016/j.energy.2018.07.019.

[67] H. Zheng, J. Yuan, and L. Chen, "Short-Term Load Forecasting Using EMD-LSTM Neural Networks with a Xgboost Algorithm for Feature Importance Evaluation," *Energies 2017, Vol. 10, Page 1168*, vol. 10, no. 8, p. 1168, Aug. 2017, doi: 10.3390/EN10081168.

[68] P. Li and J. S. Zhang, "A New Hybrid Method for China' s Energy Supply Security Forecasting Based on ARIMA and XGBoost," *Energies 2018, Vol. 11, Page 1687*, vol. 11, no. 7, p. 1687, Jun. 2018, doi: 10.3390/EN11071687.

[69] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognit*, vol. 91, pp. 216–231, Jul. 2019, doi: 10.1016/J.PATCOG.2019.02.023.

[70] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," *Proceedings - International Conference on Pattern Recognition*, pp. 3121–3124, 2010, doi: 10.1109/ICPR.2010.764.

[71] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognit*, vol. 30, no. 7, pp. 1145–1159, Jul. 1997, doi: 10.1016/S0031-3203(96)00142-2.

[72] L. A. Jeni, J. F. Cohn, and F. de La Torre, "Facing imbalanced data - Recommendations for the use of performance metrics," *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*, pp. 245–251, 2013, doi: 10.1109/ACII.2013.47.

[73] D. M. W. Powers and Ailab, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," Oct. 2020, doi: 10.48550/arxiv.2010.16061.

[74] V. García, R. A. Mollineda, and J. S. Sánchez, "Index of balanced accuracy: A performance measure for skewed class distributions," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5524 LNCS, pp. 441–448, 2009, doi: 10.1007/978-3-642-02172-5_57/COVER.

[75] C. Ferri, J. Hernández-Orallo, and R. Modroiu, "An experimental comparison of performance measures for classification," *Pattern Recognit Lett*, vol. 30, no. 1, pp. 27–38, Jan. 2009, doi: 10.1016/J.PATREC.2008.08.010.

[76] J. Muschelli, "ROC and AUC with a Binary Predictor: a Potentially Misleading Metric," 2020, Accessed: Aug. 30, 2022. [Online]. Available: www.epeter-stats.de/roc-curves-and-ties/,

[77] S. D. Watson, K. J. Lomas, and R. A. Buswell, "Decarbonising domestic heating: What is the peak GB demand?," *Energy Policy*, vol. 126, pp. 533–544, Mar. 2019, doi: 10.1016/J.ENPOL.2018.11.001.

[78] S. H. Hong, T. Oreszczyn, and I. Ridley, "The impact of energy efficient refurbishment on the space heating fuel consumption in English dwellings," *Energy Build*, vol. 38, no. 10, pp. 1171–1181, Oct. 2006, doi: 10.1016/J.ENBUILD.2006.01.007.

[79] Terry Therneau, Beth Atkinson, and Brian Ripley, "Package 'rpart,'" 2022, Accessed: Aug. 31, 2022. [Online]. Available: https://cran.r-project.org/package=rpart

[80] N. L. Afanador, A. Smolinska, T. N. Tran, and L. Blanchet, "Unsupervised random forest: a tutorial with case studies," *J Chemom*, vol. 30, no. 5, pp. 232–241, May 2016, doi: 10.1002/CEM.2790.

[81] R. Gomes, M. Ahsan, and A. Denton, "Random Forest Classifier in SDN Framework for User-Based Indoor Localization," *IEEE International Conference on Electro Information Technology*, vol. 2018-May, pp. 537–542, Oct. 2018, doi: 10.1109/EIT.2018.8500111.

[82] "XGBoost Parameters — xgboost 1.6.2 documentation." https://xgboost.readthedocs.io/en/stable/parameter.html (accessed Aug. 31, 2022).