Data-driven modelling of energy demand response behaviour based on a large-scale residential trial

Antonopoulos, Ioannis; Robu, Valentin; Couraud, Benoit; Flynn, David

DOI
10.1016/j.egyai.2021.100071

Publication date
2021

Document Version
Final published version

Published in
Energy and AI

Energy and AI 4 (2021) 100071

Contents lists available at ScienceDirect

Energy and AI

journal homepage: www.elsevier.com/locate/egyai

Data-driven modelling of energy demand response behaviour based on a large-scale residential trial

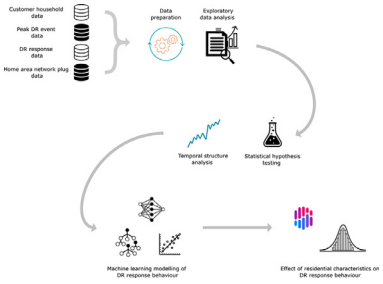
Ioannis Antonopoulos[a,*], Valentin Robu[a,b,c], Benoit Couraud[a], David Flynn[a]

[a] School of Engineering and Physical Sciences, Earl Mountbatten Building, Heriot-Watt University, EH14 4AS Edinburgh, UK
[b] CWI, National Research Center for Mathematics and Computer Science, Amsterdam 1098 XG, the Netherlands
[c] Algorithmics Group, EEMCS, Delft University of Technology, 2628 XE Delft, The Netherlands

HIGHLIGHTS

- An in-depth analysis of the response behaviour of a large number of households.
- Analysis of the temporal structure of response behaviour.
- Data-driven modelling of DR response based on household and appliance data.
- Study of the effects of important households' characteristics on response behaviour.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 21 January 2021
Received in revised form 24 March 2021
Accepted 28 March 2021
Available online 8 April 2021

Keywords:
Artificial intelligence
Machine learning
Artificial neural networks
Ensemble methods
Demand response
Residential response behaviour
Power systems

ABSTRACT

Recent years have seen an increasing interest in Demand Response (DR), as a means to satisfy the growing flexibility needs of modern power grids. This increased flexibility is required due to the growing proportion of intermittent renewable energy generation into the energy mix, and increasing complexity in demand profiles from the electrification of transport networks. Currently, less than 2% of the global potential for demand-side flexibility is currently utilised, but a more widespread adoption of residential consumers as flexibility resources can lead to substantially higher utilisation of the demand-side flexibility potential. In order to achieve this target, acquiring a better understanding of how residential DR participants respond in DR events is essential – and recent advances in novel machine learning and statistical AI provide promising tools to address this challenge. This study provides an in-depth analysis of how residential customers have responded in incentive-based DR, utilising household-related data from a large-scale, real-world trial: the Smart Grid, Smart City (SGSC) project. Using a number of different machine learning approaches, we model the relationship between a household's response and household-related features. Moreover, we examine the potential effects of households' features on the residential response behaviour, and highlight a number of key insights which raise questions about the reported level of consumers' engagement in DR schemes, and the motivation for different customers' response level. Finally, we explore the temporal structure of the response – and although we found no supporting evidence of DR responders learning over time for the available data from this trial, the proposed methodologies could be used for longer-term longitudinal DR studies. Our study concludes with a broader discussion of our findings and potential paths for future research in this emerging area.

* Corresponding author at: School of Engineering and Physical Sciences, Earl Mountbatten Building, Heriot-Watt University, EH14 4AS Edinburgh, UK.
E-mail addresses: ia46@hw.ac.uk (I. Antonopoulos), v.robu@cwi.nl (V. Robu), b.couraud@hw.ac.uk (B. Couraud), d.flynn@hw.ac.uk (D. Flynn).

https://doi.org/10.1016/j.egyai.2021.100071

1. Introduction

The increasing proportion of renewable energy resources and the growing adoption of new variable load types (e.g. Electric Vehicles) in the energy mix poses new challenges to electricity grids [1]. The effective operation of power systems demands them to function within a range of specific acceptable values for frequency and voltage. This requires the electricity demand and supply to be balanced at all times in power systems. Traditionally system operators have relied solely on solutions such as expensive network reinforcements or keeping conventional power plants open, and spinning reserve to provide reserve capacity.

An alternative approach for addressing these challenges is electricity demand response (DR) [2]. DR is regarded as a potential tool that can provide the necessary flexibility for mitigating the intermittency of renewable energy generation and the lower predictability of future loads [1]. DR solutions allow grid operators to maintain the power grid generation and load balance at a low cost, while avoiding or delaying the need for costly reinforcements of the power networks, or investing in additional back-up generation.

According to a 2020 IEA report [3], less than 2% of the global potential for demand-side flexibility is currently being utilised. The predominant sources of demand-side flexibility are industrial thermal loads and processes, thermal comfort in buildings (both residential and non-residential), charging of electric vehicles and on-site generation and energy storage [3]. This demand-side flexibility can be offered by Virtual Power Plants (VPPs), demand response providers and prosumers [2]. Demand response has been identified as a key component for providing this needed flexibility across the research literature. Many studies have explored the effectiveness and performance of demand response schemes, and they have provided evidence to support that demand response schemes (e.g. dynamic pricing schemes) can be used to provide the necessary flexibility [4–7].

The demand response base is relatively low and the worldwide inventory of flexible assets in the residential, commercial and industrial sectors needs to grow multiple times higher than it is today [3]. For the widespread adoption of demand response programmes there is the need for participation in DR from distributed residential consumers, which have a huge demand response potential. However, the diversified, heterogeneous and distributed nature of the residential DR assets makes it more challenging for service providers (i.e. demand-side aggregators) to participate in the electricity markets with demand-side aggregated resources [8].

Accurately estimating the response behaviour of DR participants and understanding the drivers behind their response is essential for addressing these challenges. An accurate forecast of response behaviour in DR schemes can decrease the uncertainty related to the available flexibility, and is directly related to the reliability of the services provided to the electricity grids. An enhanced understanding of customers' response behaviour can aid researchers, retailers and system operators to design more accurate and precise pricing or incentive mechanisms for DR programmes. The design of a DR mechanism not only affects the profitability of the participating entities, but also the success of the DR scheme. Energy big data analysis and other data-driven approaches can be a powerful tool for studying the consumers' response behaviour [9].

In this work, we base our analysis on datasets from the **Smart Grid Smart City (SGSC)** trial project data [10]. The SGSC project is one of the world's largest residential DR trials, running for almost 2 years. During the duration of the trial project there were numerous product bundles offered to the participating consumers [11]. The product bundles were a combination of pricing/incentive schemes (e.g. dynamic peak pricing, dynamic peak rebate, seasonal time-of-use, etc.) and usage feedback technologies. For a more in-depth description of the SGSC trial project the reader can refer to the project's reports and the work of Motlagh et al. [12].

The focus of this work, within the larger SGSC trial, is on the incentive-based dynamic peak rebate (DPR) scheme provided by the network trial partner. In this programme, participating customers were encouraged to reduce their electricity consumption by receiving rebate incentives. These customers were notified of the DPR events via SMS notifications 24 hours in advance of the event to request their participation [11]. All the participants in this DR scheme were residential/domestic customers.

1.1. Contributions and study outline

Our work provides new and complimentary insights, which can be used to augment prior research and brings a more complete picture of demand response behaviour, and the drivers behind it. Most studies reported in existing literature use household load data (energy consumption data at different points in time) to predict response behaviour (reduction of consumption as a response to a signal) of households, and do not consider households' characteristics. In this study, we use household-related features such as the type of heating system, internet connection, and air-conditioning units to assess their response behaviour. The main contributions of this work can be summarised as:

- In-depth analysis of the response behaviour of a large number of households based on household and appliance data, from a large real-world trial.
- Prediction of the households' response behaviour based only on household-related data, with accuracy in line with expectations — given results from previous studies on prediction of response behaviour focused on load data.
- Analysis of the temporal structure of the response behaviour and the responders' ability to learn over time.
- A detailed and interpretable study of the effects of the various households characteristics and appliances used, and their influence on energy reductions achieved.

Moreover, our research tries to promote more transparent and reproducible research by using public demand-response data and making available online the code needed to reproduce the exact results as a public GitHub repository[1]. Through this study we hope to improve the modelling and the understanding of DR response behaviour, for better planning and decision-making in the demand response domain.

The remainder of this paper is structured as follows. We start by presenting an overview of the data-driven methods utilised in this paper and a literature review of the data-driven approaches used for the study of demand response behaviour in Section 2. That way we provide the context for our approach and we position our work to the existing literature. We then perform modelling and analysis of the response behaviour, by first exploring the available response-related datasets. The first step is focused on "cleaning" the datasets and identifying potential characteristics — as well as issues with the data —, utilising existing state-of-the-art exploratory data analysis and statistical tests in Section 3. We also explored the possibility of DPR responders learning over time, and searched for temporal patterns of the response. Subsequently we applied machine/statistical learning techniques to model the relationships between the households' response and household-related features in Section 4. Having modelled the relationship between the households' features and how their response to DR events, we then try to pinpoint the important features of this relationship and we also study and discuss the effect of these features on the households' response under the DR scheme. Section 5 concludes this work with a discussion and potential extensions of this study.

[1] See the *code availability section* at the end of this study for the URL.

2. Data-driven approaches for demand response behaviour

In principle, the problem of modelling the response of consumers under DR schemes is a complex and dynamic problem based on both external and internal (to each consumer) factors. Data-driven methods offer a means for identifying patterns, modelling, and acquiring insights from the available datasets [9].

2.1. Overview of key data-driven techniques

The multitude of data-driven methods can be grouped according to their application. There are **exploratory data analysis** tools where, through the use of descriptive statistics and data visualisation, we try to study the structure of the data, variables' distribution, and the interdependencies within the data sets [13]. Data exploration is used as an initial step acting as a guide for the subsequent treatments to the data.

Data preparation is the step of pre-processing the raw data for denoising, identifying discrepancies in the data and potential systematic errors. In the context of demand response, among others, that would mean pinpointing "faulty" smart meter readings, baseline estimations, and potential presence of extreme compensation values. There is a need to pre-process data to identify missing entries from smart meters readings, customers' surveys and in general from collected data, to avoid low quality data becoming inputs to models resulting to erroneous estimates and insights. Resolving the challenge of missing data can be achieved by employing complete-case analysis or imputation techniques (ranging from simple mean imputation to multiple imputation with Markov Chain Monte Carlo [14]). In missing data analysis it is paramount to understand the underlying missing data mechanism [15], e.g. whether the data are missing, or not, at random. De-noising the data entails pinpointing and filtering out errors in the data and/or outliers. Data preparation is an important process of the data analysis pipeline; real-world data is impure and noisy, and errors and biases of raw data can propagate to subsequent steps of analysis and modelling, with a great effect on the results of data-driven approaches [16].

Furthermore, **data mining** can be used for clustering consumers behaviour to provide suggestions of consumers for DR schemes, selection of participants in DR events, and design of DR programmes. Frequently used categorisation techniques include both supervised classification and unsupervised clustering algorithms. Methods such as k-means and self-organising maps have been widely used in the context of demand response and for the study of response behaviour. In demand response, where a significant proportion of data is time-series, it is essential to use tools from **time series analysis** to discover potential underlying temporal structure. These tools include autocorrelation, identifying trends and seasonality patterns. For example, autocorrelation (or serial correlation) is a measure of linear association between lagged values of a sequence (e.g. time series). i.e. a lag k autocorrelation is the correlation between values that are k time periods apart.

Another important application is also the modelling of data and estimation of variables of interest. In DR, that would be prediction of energy demand to provide information to support energy producers in the accurate planning for energy production, estimation of load curtailment, reliability of DR participants, etc.

Widely used techniques for practical **data-driven modelling** of data are ensemble methods and artificial neural networks [17]. The most popular ensemble methods are gradient boosting and random trees. Gradient boosting [18] is an ensemble of weak predictive models (typically decision trees) and it is an additive, forward stage-wise boosting model. This means that the algorithm sequentially adds new decision trees to the model without altering the parameters and coefficients of the decision trees already added [18]. Moreover, this type of models allow for the optimization of arbitrary differentiable loss functions [19].

Random forest is also an ensemble method based on decision trees, and it operates by constructing a multiple of independent decision trees that are trained independently on a random subset of data [20]. To grow each tree, the inputs (or combinations of them) at each node are selected in a random fashion [20]. Contrary to the gradient boosting algorithm (where the weak models are built sequentially), random forests are an averaging ensemble method; meaning that they build weak decision trees independently and then they take the predictions' average of these weak estimators.

Artificial neural networks are computational models inspired by, albeit not identical to, biological nervous systems. The two basic architectures of artificial neural networks are the feed-forward and the recurrent architecture [21]. A dense neural network is a feed-forward network where all its layers are fully connected. More specifically, each unit in a dense (fully-connected) layer is connected to all the units in the subsequent layer [22]. Dense neural networks (and feed-forward neural networks in general) can be thought as universal function approximators which are constructed to achieve statistical generalisation [22].

For domains like demand response, which have not been extensively studied and evaluated, there is the need to understand how the models attained their results and not focus only on their predictive performance [23]. For that goal, there is a need for **interpreting** data-driven models. Two popular model-agnostic interpretation methods are the (SHapley Additive exPlanations) SHAP framework [24] and the permutation feature importance technique [20,25]. We have employed these two approaches because prior research has shown that, in cases where the input variables are categorical (like in this study), the measures of variable importance and Gini importance can be biased towards categorical features with more categories [26]. These two frameworks approach the problem of feature importance from different angles.

In principle, the permutation feature importance algorithm assigns importance to the various input variables by taking into account the deterioration in model performance, whereas SHAP is based on the features' contribution to the model predictions. In more detail, the model-agnostic **permutation feature importance** algorithm calculates the importance of a feature by measuring the increase/decrease in the model's error metrics when the values of this feature are randomly shuffled [25]. This permutation breaks the relationship between the feature and the target variable, thus a model with higher error metrics (positive importance) is indicative of the model's dependence on the feature. On the other hand, negative importance (lower error metrics after permutation) means that the feature is not important for the model [20].

The **SHAP** approach is a framework for interpreting the models' predictions, by assigning to each feature an importance value for a particular prediction [24]. The SHAP framework explains the prediction of a data instance x by computing the contribution of each feature [24]. The contribution of a feature is calculated by computing Shapley values from coalitional game theory [27]. Under this paradigm, the players in a coalition are the feature values of data instance x and the Shapley values denote how the pay-out (in this case the prediction) is distributed among the features in a fair manner.

2.2. Related work on data-driven methods for DR response behaviour

The subject of the analysis and modelling of customers' response behaviour has been broadly studied across the DR literature (see Antonopoulos et al. [28] for a full review). There are numerous papers where data-driven techniques have been used to model and predict customers' response to DR signals. Zhou et al. [29] estimate the reductions in electricity consumption during demand response time-windows, by incorporating latent variables (i.e. not observable variables) in statistical forecasting models — i.e. Ordinary least squares, k Nearest Neighbours, Support vector, and decision tree regression models. The latent features are constructed from the consumer's consumption data by applying sequentially a Conditional Gaussian Mixture model and a Hidden Markov model. Liu et al. [8] train a Long Short-term Memory (LSTM) neural network on simulated consumers' response data. The response data were generated using a response function which is quadratic to the customers' received incentives.

Paterakis et al. [30] approach the problem of the prediction of customers' response under dynamic DR price signals, by using a hybrid load forecasting model. The time-series load data are decomposed to various frequency signals, by applying wavelet transform, which are then fed as inputs to dense neural networks. The outputs from the neural networks are aggregated to create the final response forecast. Holtschneider and Erlich [31] estimate the response behaviour to DR price signals by applying a dense neural network. To tackle the issue of sparse training data for predicting response behaviour — there are only a few DR events per year —, there is work where they have employed non-parametric approaches i.e. an ensemble of k-Nearest Neighbours regression models [32], or transfer learning-based approaches [33]. In the work of Cai et al. [33] they attempt to predict the customers' response behaviour in incentive-based DR by using load data from DR participants with similar load consumption behaviour.

The input data to the aforementioned data-driven algorithms have been found to be numerous.They are features engineered from the consumers' load data i.e. maximum and minimum daily load, mean consumption [8,32], as well as the actual data of consumers' electricity consumption [29,30,33]. Quite often time-related features are included in the model i.e. week, month, hour of the DR event, seasonal scores [8,32], as well as environmental factors i.e. ambient air temperature [29,33] and humidity [33]. The target variable for DR response prediction has mainly been the electricity consumption of consumers [8,29,30]. Cai et al. [33] forecast the curtailed load (difference of the actual electricity consumption from the estimated baseline consumption), and Kang and Lee [32] the proportion of realised load reduction to the requested load reduction. The training data of the learning algorithms has primarily been simulated or is semi-synthetic data [8,29,30], but there are also cases where real-world DR data have been used [32,33].

Moreover, clustering the main DR participants attributes is a typical approach for analysing the various response behaviours. The literature has mainly based their analysis on the customers' load data for the analysis of the response behaviour. Motlagh et al. [12] have focused on knowledge extraction from load profile data by comparing customers' consumption behaviours under DR with the ones of the control group. They use Principal Component analysis to extract the principal components from half-hourly consumption data and then they cluster them, based on these principals components, by using self-organising maps. Other clustering algorithms like k-means [34], expectation maximisation [34], and finite mixture-based clustering [35], have been employed for understanding the DR participants behavioural use.

Other widely applied AI approaches used to study the behaviour of demand response are mechanism design and cooperative game theory — which are often modelling, not data-driven. These methods have been employed for the design of demand response contracts, such as in [36–38], demand response incentive mechanisms [39,40], demand-side and consumer coalitions [41–43], or coordinating specific devices, such as electric vehicles [44,45].

It is noted that, while the previous papers have provided valuable insights and advancements they tend to focus on the analysis of response using primarily load data. By contrast, the purpose of this work is to study the modelling and analysis of response based on household and device-related features. With specific emphasis on studying the effects of those features on the DR response behaviour.

3. Individual datasets exploration and analysis

In this study we use data from the Smart Grid, Smart City trial project. The SGSC project is one of the largest commercial scale smart grid technology trial projects globally [12]. It includes a wide spectrum of trial data, ranging from household and device-related data to EV utilisation. In our research the SGSC trial datasets utilised are the following:

- Customer household dataset
- Peak events dataset
- Peak event response dataset
- Home Area Network Plug Readings (HANPR) dataset

The household dataset includes data related to the participating households including the type of product they were involved with, energy and gas usage tier, type of house, type of devices in the household among others. The Home area plug readings dataset has the energy usage readings of various household devices, and can offer information, in a more granular level, of what type of electricity devices the customers utilise. The peak events dataset contains temporal details regarding the DPR and DPP events, whereas the peak event response set has the actual consumption, baseline estimation and rebate amount data. Before applying machine learning techniques for modelling DR response behaviour, we use exploratory data analysis and data exploration tools to pre-process the various datasets.

3.1. Household & HANPR data analysis

Initially in the household dataset we drop some features irrelevant to our analysis. That includes information on non-tariff products(e.g. lifestyle audit, feedback technology), redundant features (e.g. internal references), and in general non-informative features for how participants have responded in DR events. Although in the SGSC trial customers were offered a few tariff products, in this work the analysis is focused on the incentive-based DPR product. Therefore, we filtered out all the households which did not participate in the DPR scheme. After the filtering stage the resulting dataset did not have any missing entries.

The household dataset includes both numerical and categorical features (qualitative variables) with the majority of the features being categorical variables. When there is a mixture of both types it can be tricky to visualise the level of association between all these features. In this work, the approach we have selected is splitting the dataset in two subsets (based on the type of features) and calculating the Spearman correlation for the numerical variables and the Cramér's V[2] metric [46]. Figs. 1 and 2 visualises the level of association for numerical and categorical features respectively.

Especially in our study, where one of the main purposes of the regression models is to investigate the relationship between the input variables and the households' response for DR, multicollinearity among the input variables can lead to misleading analysis of the effects of the input variables on the DR response, as well as erroneous interpretations of the fitted models [24,47,48].

From Fig. 1 we can see that in general the numeric features are not highly correlated with each other. The only case where there is a somewhat significant correlation is among the features related to household's power generation meters (number of gross and net solar generation meters, number of general supply meters, number of miscellaneous generation meters); which is to be expected. In Fig. 2 we can identify that there are not really strong associations among the categorical features, with the exception to the following cases below:

- Between the assumed dwelling type and the actual dwelling of participating household.
- Between the air-condition type and the response to whether the participating household has air-condition.
- Among the gas-related features of the household.

In relation to the HANPR data, where there is more granular information about each household's devices, we explored whether we could engineer features related to the participants devices. That was done for the potential use of these features in the prediction of each household's response to the DPR events. Unfortunately this dataset only had relevant device data for a small proportion of the total participating households set, therefore, we have not utilised these features further in our analysis.

[2] It is measure of association between two nominal variables — based on Pearson's χ^2 statistic —, giving a value between 0 and +1. [46]

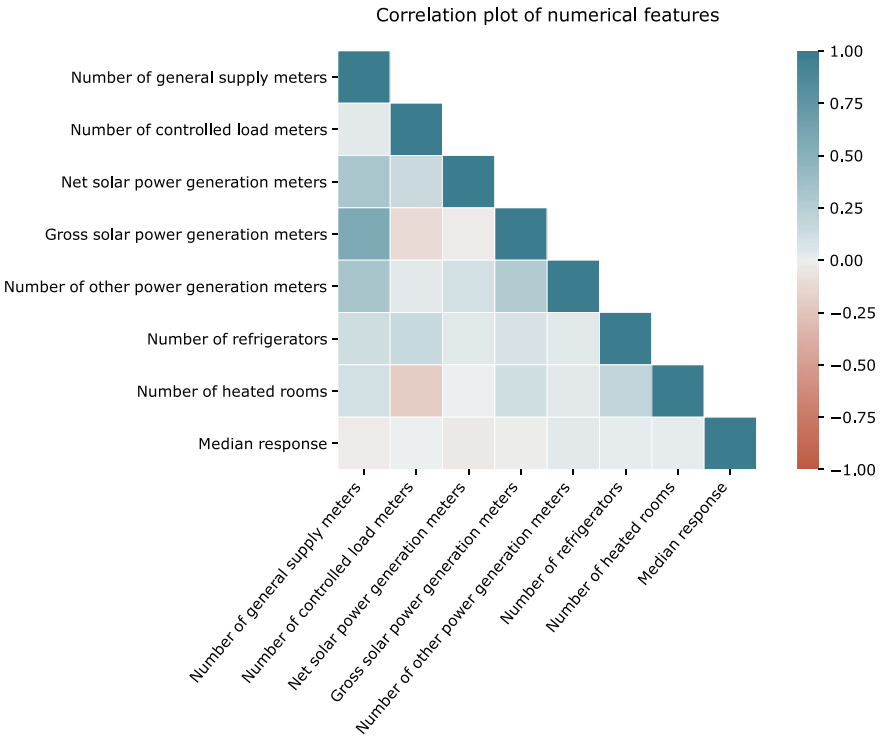


Fig. 1. Spearman's correlation among the numerical household characteristics. The colour bar indicates the value of the correlation ranging from −1 to +1.

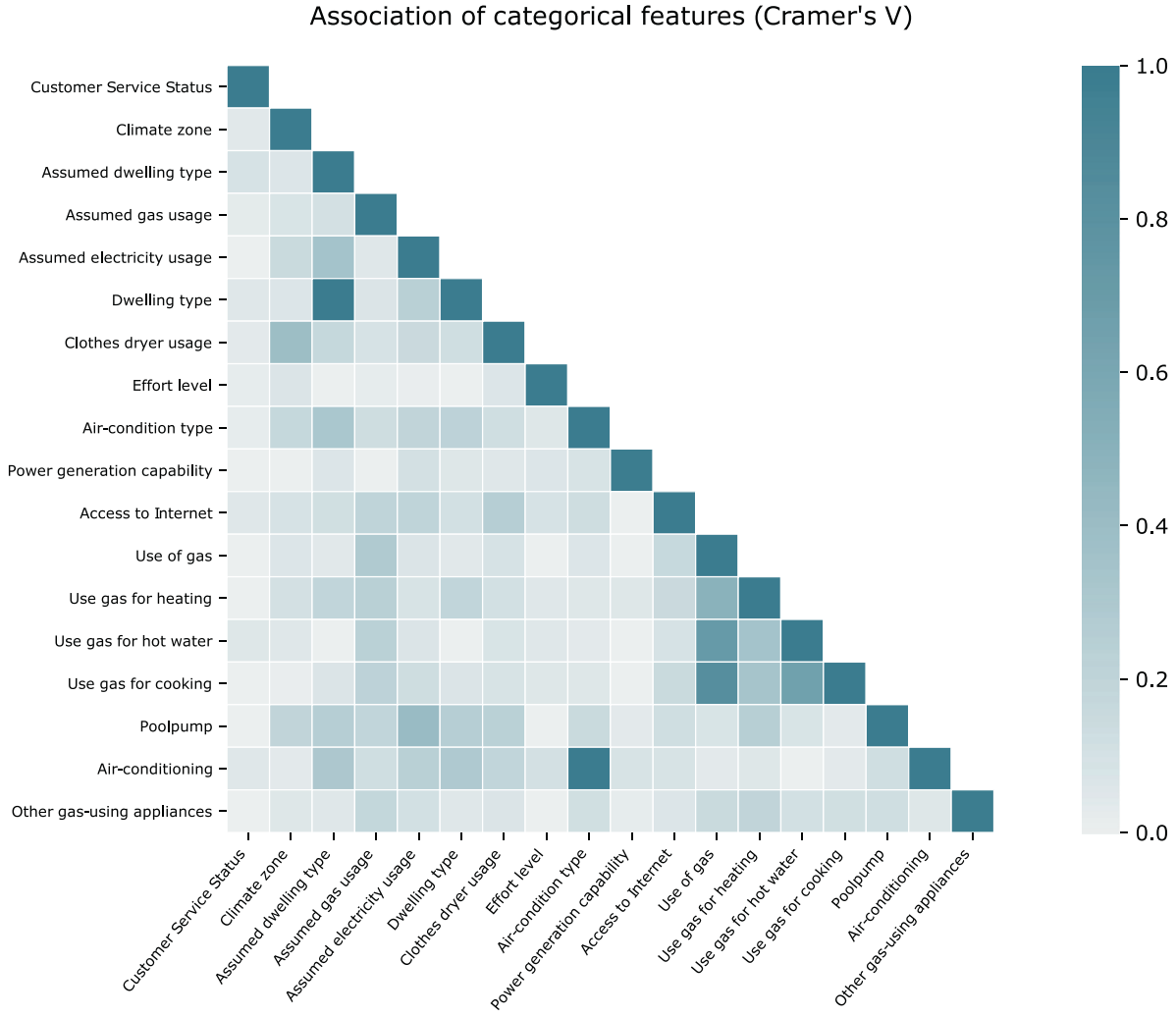
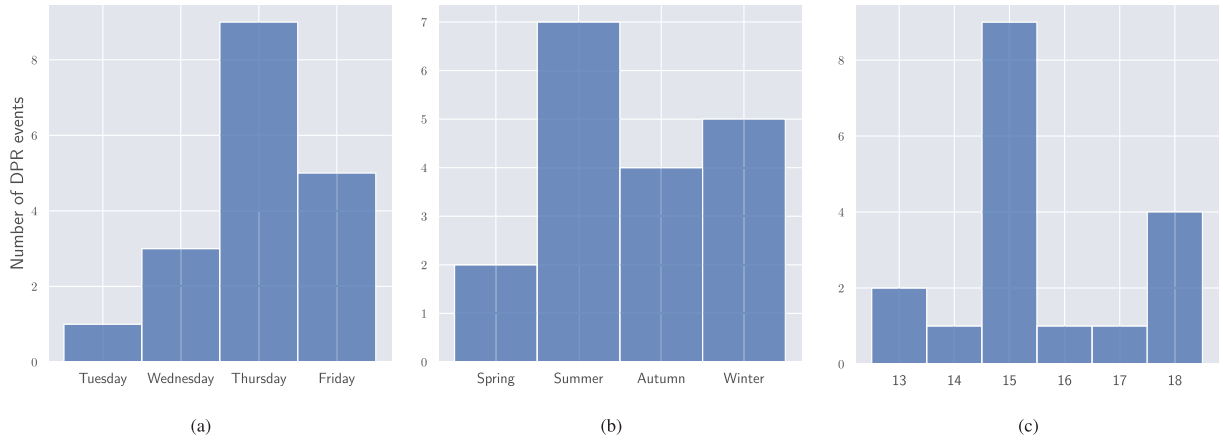


Fig. 2. Cramér's V association for the categorical household features. The colour bar indicates the value of this association measure, ranging from 0 to 1.



(a) (b) (c)

Fig. 3. Number of DPR events per day (a), per season and (b), and per hour of day (c).

3.2. Peak events & response data analysis

Among the products that were offered in the trial project there were two peak event products: dynamic peak pricing (DPP) and dynamic peak rebate (DPR). In this study, the focus of modelling and analysis is the incentive-based DPR programme. In the DPR scheme the participating households were offered a rebate amount (constant at $AUD 4.5/kWh$) to decrease their electricity consumption compared to their baseline estimation.

Across the running time frame of the SGSC trial there were 18 DPR events, as it is illustrated in Fig. 9 of the Appendix A. The duration of the DPR events ranges from two to four hours, and they have only happened during the weekdays (mainly Thursday and Friday), primarily during the summer and winter months, as well as mainly in the late afternoon hours, as shown in Fig. 3. Since this is a dataset of domestic consumers from Australia, it is natural to see the peak events occur in the summer in early afternoon, when there is maximal use of air conditioning units.

Moving on the DPR response data, for the rebate programme we have complete data for only 13 out of the total of 18 DPR events (found in the peak event dataset). This means that we have the actual consumption and the baseline estimation for all participants, as well as the rebate amount, for 13 peak events. At this point, we should note to the reader that when we use the term **response** to a DR event we are referring to the **percentage difference** between the **actual** electricity consumption of the household and its estimated **baseline** consumption, for the duration of the specific DR event — as described in Eq. (1).

$$r_{i,j} = \frac{load_{i,j} - base_{i,j}}{base_{i,j}} \times 100 \tag{1}$$

Where $r_{i,j}$ is the response of household i for the DR event j. Respectively, $load_{i,j}$ is the actual consumption, and $base_{i,j}$ is the baseline consumption (estimated by the service provider) of household i for the duration of the DR event j. This metric was selected with the reasoning that it has the advantage of being comparable across different customers, contracts, and DR schemes. But, at the same time there are limitations; i.e. errors of it constituent parts (errors in baseline estimation) can propagate to the response variable. Due to the fact that baseline estimation is one of the basic issues for demand response schemes [49,50], we investigate any potentially faulty estimations of baseline and we filter out these data points.

First, given that a typical household consumes in average more than 0.45 kWh per 2 h, we exclude from our analysis all the cases where the baseline estimation is lower than 0.01 kWh. We can infer that households consuming less than this electricity consumption indicate probable faulty values in the baseline estimation, or that the households are not really used by their owners and therefore there is no reason for including them in our response analysis.

In Table 1 we provide the descriptive statistics of the various customers' actual and baseline consumption across the DPR events, as well as of their received rebates and their response for the DPR events. The descriptive statistics include the mean value, the standard deviation (std), minimum and maximum values, and the quartiles (25%, 50%, 75%)

We can see that the average response for the DPR scheme is positive (34.25%), meaning that on average a participating household will not decrease their consumption during a peak (DPR) event but actually increase it by around a third. We can also see that the standard deviation (std) of how customers have responded in the DPR events is exceptionally high (≈ 800%). This could be partly attributed to outliers and potential issues with baseline estimation. Looking at the 3rd quartile (75%) value of response we can see that it is positive. That means that quite often (> 25% of the cases) households have not actually decreased their consumption during a DPR event. The extremely high maximum values and standard deviation of responses indicate an issue with outliers in the data.

We filtered out the DPR events which show irregular distribution patterns, and to address the issue found with the outliers in the remaining DPR events we employed unsupervised techniques for outlier/anomaly detection. The outlier detection is implemented on each DPR event, and not on each household. Mainly because the baseline estimation can change through time, and also because the number of DPR events per households is really small to do any meaningful outlier detection.

The algorithms that we examined for this task are the Local Outlier Factor (LOF) algorithm [51], and the Isolation Forest [52]. The applied LOF algorithm is a density-based approach which employs the k-nearest neighbours to estimate the local density of each sample, and tries to identify samples which have significantly lower density to its surrounding datapoints. The isolation forest algorithm takes a different approach to outlier detection, with explicitly isolating the anomalies instead of profiling normal samples [52]. It uses an ensemble of random splitting trees, and considers outliers those datapoints which have short average path lengths on these trees.

In this study we applied an approach similar to the algorithm proposed in Cheng et al. [53] where the two algorithms are combined to try and overcome their respective limitations. First we applied the Isolation forest and extracted an initial set of outliers, and on that initial set the LOF algorithm was applied to perform local outlier detection and get the final set of outliers. The distribution of the various households' response $r_{i,j}$ per DPR event after removing the outliers/anomalies is illustrated on Fig. 4. From Fig. 4 it is apparent that even after removing the outliers there is a high variance in how households have responded to DR signals. This result is in accordance with previous work in the literature [4], where it is presented that the average reduction DR trials is ranging between 10% and 50%. The issue

Table 1
Descriptive statistics of the consumers' actual and baseline consumption, rebate amount and response across the DPR events.

	Actual consumption (kWh)	Baseline consumption (kWh)	Rebate amount (AUD)	Response (%)
mean	3.331	5.360	14.501	34.247
std	3.620	4.889	17.148	799.552
min	0.000	0.100	0.000	−100.0000
25%	1.031	1.865	2.803	−70.637
50%	2.094	3.936	9.100	−40.411
75%	4.293	7.405	19.660	7.204
max	39.565	50.972	164.800	66218.750

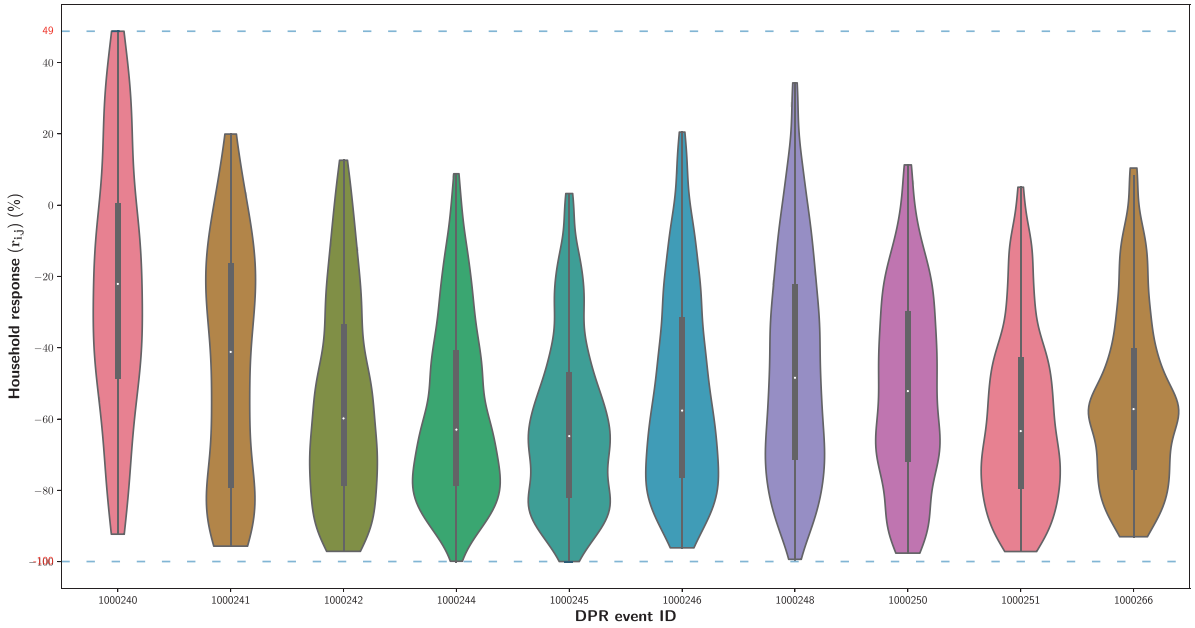


Fig. 4. Violin plots showing the probability density of households' response (as defined in Eq. (1)) for ten DPR events. The circle marker shows the median of the data and the box inside each violin plot indicates the interquartile range.

of variance in DR has been discussed also in the work of Aïd et al. [36].

A subsequent step, following the removal of the outliers from the dataset is to study the distribution of the households' responses to the various DPR events. We filter out the customers who have responded to less than five (out of total 10) DPR events. The reasoning behind this filtering of consumers is that there should be a minimum number of events where they have responded, in order for this analysis to provide any meaningful insights. Fig. 5 illustrates that the distribution of the consumers' responses seems to be similar to a *log-normal* distribution. The distribution of the responses in this case is not symmetric and appears to be positively skewed. This tendency is the same across all seasons and it is apparent that it persists in the aggregated case too. When the data follow a skewed distribution, the median is often the best measure of central tendency. That is the rationale behind selecting the median response per household (across the DPR events in which they have participated), rather than the mean as a measure of central tendency.

3.3. Temporal structure analysis

A key hypothesis to explore would be whether participants are learning to respond better over time. This hypothesis has implications on the selection of consumers for response in DR events. In the case that consumers learn over time to respond better over a sequence of DR events, then a more dynamic view on the customers' potential for DR should be considered. For example, an aggregator could select these consumers more often with the intention of a better response in the longer-term. One way to check this hypothesis is with the use of autocorrelation.

We calculated the autocorrelation function (ACF) for every household's DPR responses and the distributions of 1-lag and 2-lag autocorrelation are presented in Fig. 6. At first glance there appeared to be quite a few households that have relatively high autocorrelation values. But we need to confirm that the autocorrelations in the population from which the sample is taken are statistically different from zero, rather than these observed autocorrelations being the result of randomness from the sampling process. This can be achieved by employing the Ljung Box Q test [54], which is a portmanteau[3] statistical test. In order to reject the null hypothesis of the Ljung-Box Q test (that the data are independently distributed) the p-values must have lower value than the alpha significance level.

By checking the p-values associated with the Ljung Box Q-statistics, at the 5% significance level, we can see that only 29 households had a statistically significant 1-lag autocorrelation and 21 households had a statistically significant 2-lag autocorrelation — out of a total of 953 households. With these results, there is little evidence that there is a temporal structure in the participants' response to the peak events and that there is a learning process over time, at least for this dataset where the majority of the customers participated only in a limited set of DR events. This points to an important insight for future DR trials, that could

[3] Type of statistical hypothesis test where the null hypothesis is well specified, contrary to the alternative hypothesis which is more loosely specified [55].

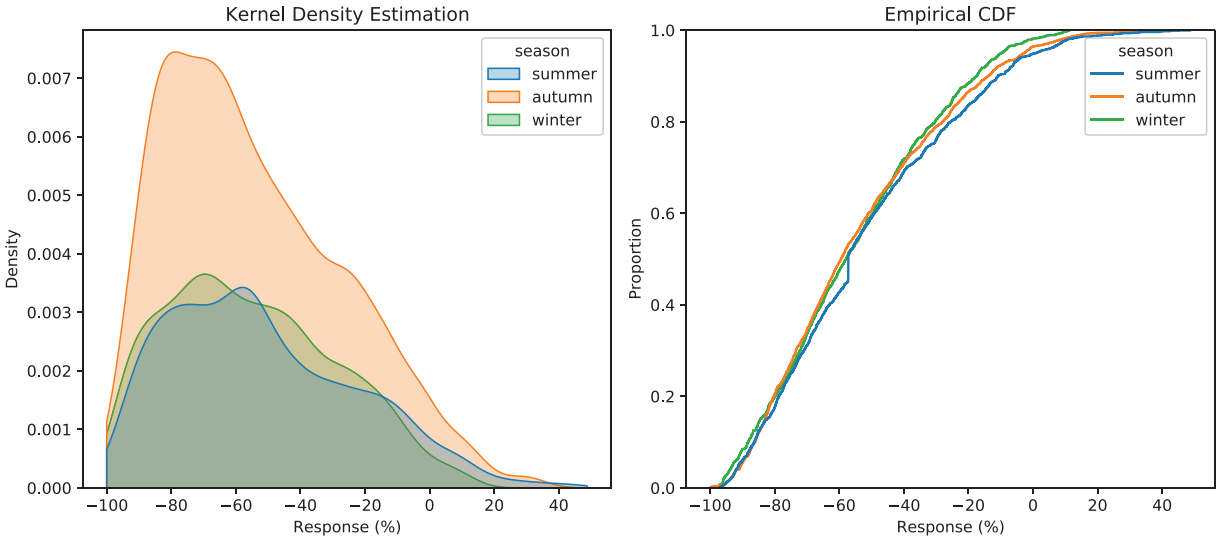


Fig. 5. Distribution plots on how households have responded per season.

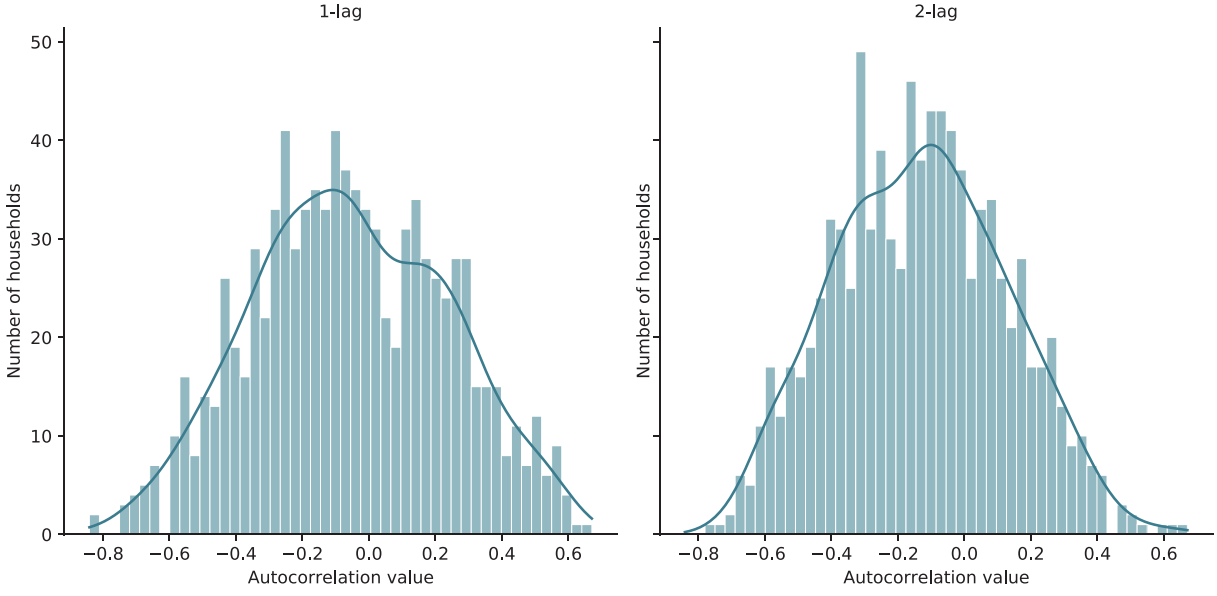


Fig. 6. Distribution plots of the 1-lag (left) and 2-lag (right) autocorrelations.

be conducted over a longer period of time, allowing better observation of temporal structure and learning.

Furthermore, by examining the response of households in Figs. 5 and 7 it can be inferred that the response does not differentiate much among the seasons. The same can be said for distinct days — DPR events only happened in three days and all of them were during the work week —, and for distinct hours and months. Their distributions can be seen in the Figs. 14–16 of the Appendix A.

4. Modelling of demand response behaviour

In this section we explore the potential relationship between the household-related features described in Section 3.1 and the median response per household for DPR DR events, and extract the important features for that purpose. As stated above (see Fig. 5) the median response per household have been selected as a target variable due to the skewed distribution of the responses, and the nature of the inputs variables (which is at the household level). The distribution of the target variable is presented in Fig. 7. For that purpose we train different types of predictive models on the household data set, and the models used in this work are the following:

- Linear model
- Gradient Boosting Regression
- Random Forest regression
- Dense Neural Network

Initially, we split the features' data in training, validation and test set and then we pre-process the data. Our dataset is predominantly composed of categorical features (see Section 3.1). As a result, the pre-processing of the dataset mainly involves converting the categorical features to numerical, using an embedding system (e.g. one-hot encoding or ordered target encoding). In general, the numerical features were not pre-processed because they are of similar scale and distribution. Moreover, in our case there were not any missing data, so there was no imputation step involved. In the subsequent paragraphs we describe more in detail the models used, and the actual encoding system used for each model.

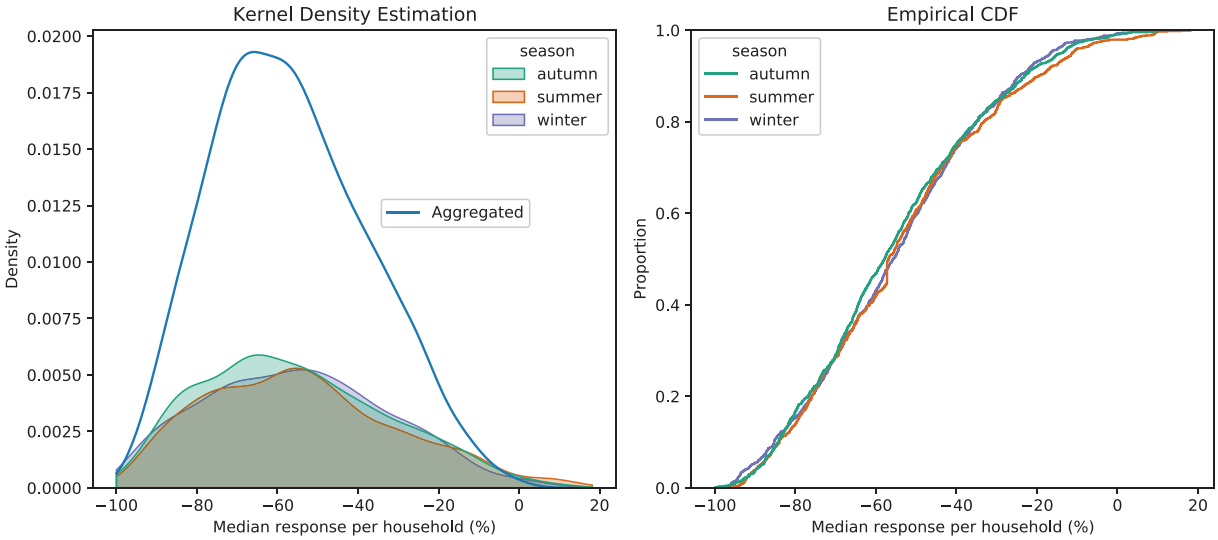


Fig. 7. Distribution plots of the median response per household and its breakdown in seasons.

In the linear model the estimation of its parameters is done using the Ordinary Least Squares (OLS) method, which is a widely used and common approach. This kind of linear model offers good interpretability of the potential relationship between the predictor variables and the target variable [56], and serves as a benchmark for the comparison of the other learning algorithms. In this work we utilised the statsmodels implementation [57] of an OLS linear regression, with one-hot encoding of the categorical variables.

In this study, the gradient boosting on decision trees algorithm is based on the CatBoost implementation [58]. This implementation includes a permutation-driven alternative to the classic gradient boosting algorithm, and has a built-in embedding approach for categorical features based on target statistics [59]. The number of trees used in the final model are selected, based on the iteration that outputs the optimal evaluation metric in the validation set. For the training of the Random Forest model we utilised the scikit-learn [60] implementation with its default parameters. The pre-processing of the dataset was done in a similar fashion to the gradient boosting case.

Here, we constructed a two-hidden layer dense neural network using the Keras API which runs on top of the TensorFlow platform [61]. The network's hyperparameters — in our case the number of units for each layer and the learning rate of the Adam optimiser [62] — were tuned based on the Keras Tuner library [63], with the use of the random search method [64]. We also experimented with "deeper" architectures (higher number of layers) but there was no gain in the predictive performance of the model, and therefore were not selected as the final model. According to the survey of Hancock and Khoshgoftaar [65] the most common encoding for categorical variables in neural networks is One-Hot encoding. In our case we have used the CatBoost encoder (as in the earlier models) for consistency reasons and to make the results more comparable.

The various machine learning models were trained and validated on 10 different splits of the dataset (in training, validation and test sets). That was done in order to evaluate their average performance and give more informed values of their test error metrics. The selected metrics are *root mean squared error (RMSE), mean average error (MAE), mean average percentage error (MAPE)* and R^2, and were evaluated on the test set of the data. The results are presented in Table 2. For a more detailed description of the models' implementation and the exact values of the optimal hyperparameters, the reader can always refer to the Jupyter notebooks/codebase in the GitHub repository as found in Section 5.

Based on the combination of the models performance on the different error metrics, the Gradient boosting algorithm seems to produce the best results in the test set, followed by the linear model. In general, the results across all these four models are quite similar (both for linear and non-linear estimators), and comparable to the existing literature. For example, the MAPE metric of models for individual customers' response is in the range of 41–45% [32,33]. In all cases the R^2 is quite low, and therefore we can infer that this household features are not very informative predictors of the median response. But, even though the input variables explain one a small percentage of the variance in the model, that does not necessarily mean that they need to be discarded as useless. In this work, the purpose is to explore the relationship among the features and their importance, and not high prediction accuracy. When the independent variables are statistically significant, we can still draw important conclusions about the relationships between the variables.

Examining the results' summary of the linear model, we can see that the in-sample R^2 ≈10%, and primarily that the p-value of the F-statistic is quite low. In the multiple regression setting this fact means that there is at least one statistically significant feature, and therefore there is a relationship between the household features (at least of a subset of them) and the household's median response. Moreover, examining the p-values of the t-test for individuals features we can see that, at the 95% significance level, there are the following statistically significant features:

- The level of a household's dryer usage
- If a household has internet access
- If the household is using gas oven
- If the household is using gas for hot water
- If the household has Air-condition
- The number of a household's controlled loads
- The number of refrigerators in a household
- Whether the household stayed on the trial for the duration

The above are supporting evidence that there is a statistically significant relationship between the input variables and a household's median response, albeit not a strong one. Therefore we extend further our analysis by exploring important features and discussing the contributions of these important features on the customers' response, in an incentive-based DR scheme.

4.1. Identification of important features and discussion of their effects on response behaviour

Given that there is a relationship between the customer households' features and the response of a household, an interesting subsequent step is to examine which features are important for the response behaviour of the DR consumers. Identifying which features are important cannot

Table 2
Average values of evaluation metrics (in the test set) for various learning models.

	RMSE	MAE	MAPE	R^2
Linear model	18.769 ± 0.794	15.284 ± 0.625	0.465 ± 0.099	0.009 ± 0.027
Gradient Boosting regression	18.711 ± 0.714	15.246 ± 0.631	0.473 ± 0.101	0.015 ± 0.018
Random Forest regression	19.721 ± 1.611	16.107 ± 1.389	0.484 ± 0.101	−0.108 ± 0.240
Dense Neural Network	18.906 ± 0.641	15.469 ± 0.663	0.484 ± 0.101	−0.008 ± 0.063

only potentially lead to more accurate prediction of response behaviour, but also provide insights to the data and the models utilised.

In our study we have applied two popular model-agnostic interpretation methods to examine **features' importance**. The SHAP framework and the permutation feature importance method. These two methods have been applied to different models in order to get more decorrelated results and a less biased set of important features. The permutation importance algorithm has been applied for the linear and random forest regression models. The results for the test set appear in the Fig. 12 of the Appendix A. In the x-axis of the importance graph we have the importance of the feature (difference between the model's score metric with and without each feature's permutation), and in the y-axis we have the features ranked by their importance magnitude. The SHAP method has been applied for the gradient boosting and dense neural network models. The importance of each feature in this case is the mean of the absolute Shapley values across the test data. The sorted features by decreasing magnitude of importance are illustrated in Fig. 11 of the Appendix A.

Based on these four different models and two feature selection frameworks we have four distinct rankings of the features' importance. To overcome this diversity among them and create a global set of ranked important features we have utilised an ensemble feature ranking based on a Markov Chain rank aggregation method [66].

Understanding the drivers behind response behaviour in DR schemes and their effect on response can have wide implications on the use of demand response for the provision of demand-side flexibility services. Among others, a deeper comprehension of residential response behaviour can assist service providers with targeting households for DR, better selection strategies for responding in individual DR events, as well as with the design of more successful, fair, with high participant engagement DR schemes.

In this work, we approach the study of the effect of households' characteristics on the residential response behaviour in two ways. We examine the distribution of the residential responses across the various important households' characteristics, and we investigate the impact of these features on the Gradient boosting model's output, employing the versatile SHAP interpretable framework. The local explanations of the SHAP method for the samples of the test set result to global model insights. The Gradient Boosting regression model has been selected because it is the model with the best prediction accuracy (across the models used in this study). The intuition behind using two different methods is that this way we cross-check the results between them, and therefore can acquire insights, on the residential response behaviour in DR events, which are potentially less biased.

In more detail, based on the results of the ensemble ranking algorithm we select the features with the highest rank, and for each feature we explore how the various levels of the categorical input variables influence the households' response. How the various numerical features affect consumers response in DPR events can be seen in Fig. 8. E.g. households with high number of refrigerators tend to not follow the DR request, whereas households with a low number of refrigerators implies that the household will decrease more their consumption during DR events. Fig. 13 presents the empirical cumulative distribution function for the different levels of the most important categorical variables. The rationale for using this type of plots is that by comparing the CDFs for different categories we can estimate whether the categories influence differently the response behaviour of the DR participants. Categories of the same variable with *"steeper"* CDFs indicate that households belonging in this category tend to have higher response, and decrease more their actual electricity consumption (compared to their baseline) than households belonging to the other categories — and vice versa.

The graphs that illustrate the contributions of each feature to the Gradient boosting regression's model are Fig. 8 and Fig. 10. Each point in these graphs is the Shapley value for a specific feature and an instance of the model's output. Fig. 8 combines feature importance with feature effects. The features are sorted by their importance magnitude and the colour represents the value of the feature from low to high. The grey colour denotes a categorical variable, where there is no notion of high and low values. Although this plot illustrates information about the relationship between the values of households' characteristics and the impact on the model's prediction, for numerical variables, it is not really informative for categorical variables. In this case to see the exact form of the relationship we have to look at Fig. 10, where we plot the Shapley value of that feature with its categories for all the examples in the test dataset. On the y-axis are the SHAP value for that feature, which represents how much knowing that feature's value changes the output of the model for that sample's prediction. Also, because of the way we have defined the households' response (percentage difference between actual and baseline consumption), negative SHAP values indicate decrease in the electricity consumption and thus better response. Moreover, the points in this plot are coloured based on the level of the household's electricity usage level. If an interaction effect is present between the plotted feature and the electricity usage level it will appear as a distinct colouring pattern.

After careful examination of the aforementioned graphs we can provide the following insights on the effects of households' characteristics on the DR response behaviour. As we can see from both approaches, households that have access to the internet tend to respond better in DR events. We can infer that this feature does not have a direct effect on the DR response, but most likely it is associated with characteristics of the residents. One interesting insight that we can infer from Figs. 10 and 13 (and seems counter-intuitive) is the fact that the reported effort level taken by the households, to respond to a DR signal, does not seem to have an effect on the households' actual response. Households that own air-conditioning systems respond better than the ones which do not. The households that do not have a gas-fired oven tend to respond better than the ones that do.

Another interesting insight is that the households with medium electricity usage level tend to respond better than the ones with low or high usage levels. Existing studies on the literature have indicated that high-use households tend to respond significantly more in load reduction than low-use households [7]. This could potentially be attributed to the design of this specific DR peak scheme, where the rebate amount per energy unit decreased is constant. Moreover, the customers in warmer climate zones tend to have a better response behaviour than the ones living in areas with mild climate conditions. The residential customers with medium and high level of clothes dryer usage respond better than residential properties which do not use the dryer frequently (or at all).

We can see that consumers with medium levels of gas usage tend to be more responsive in DR peak events than the ones with low levels. It seems that households with high levels of both gas and electricity usage tend to perform worse in response behaviour. The dwelling type does

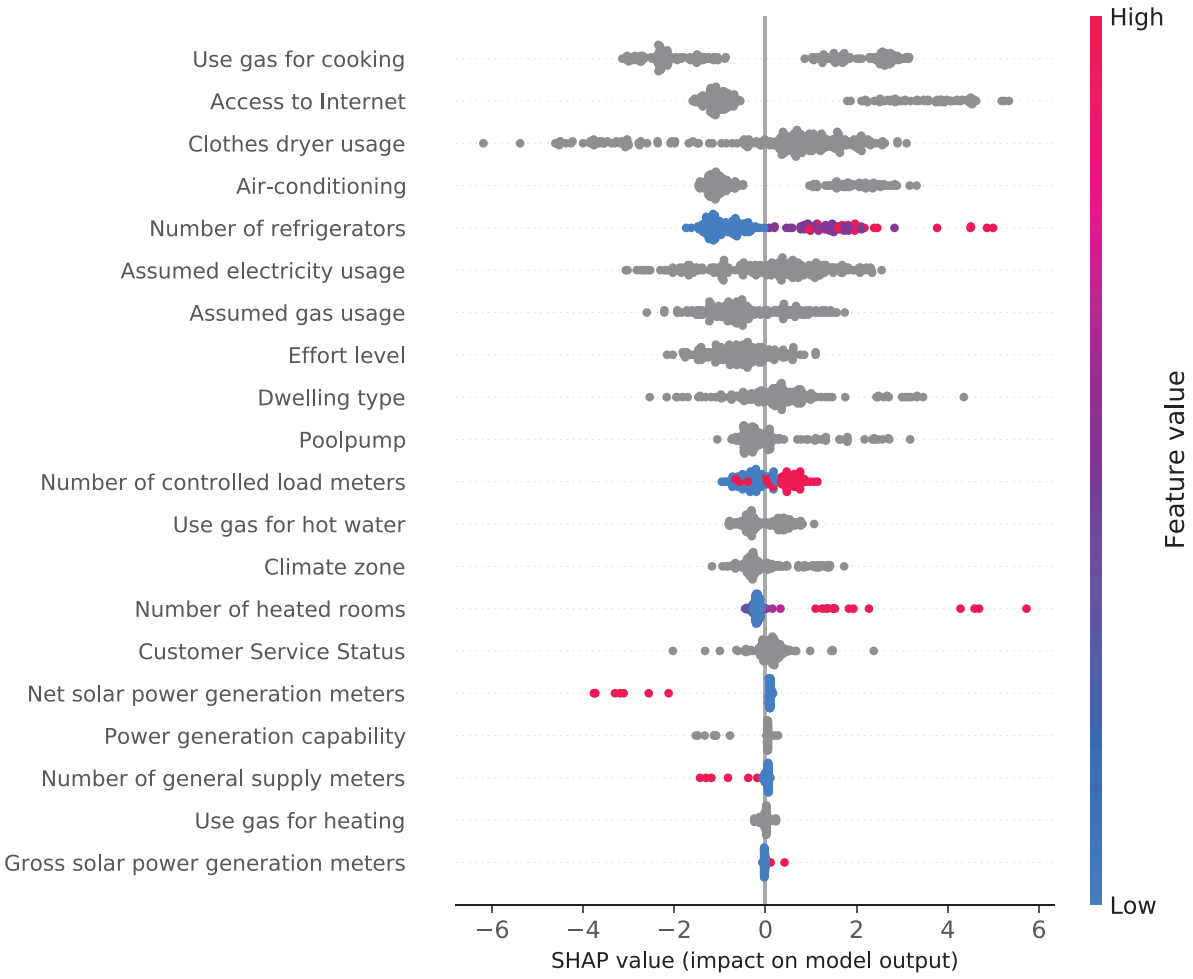


Fig. 8. SHAP values of each household characteristic (feature) for every sample of the data.

not seem to have a big effect on the response behaviour of customers. We could cautiously infer that the consumers residing in units (apartments) with medium electricity usage level tend to respond better.

5. Conclusion and further work

For the wider realisation of the demand-side flexibility potential, the widespread utilisation of residential DR resources is essential. However, residential response behaviour poses certain challenges. Accurate modelling of response behaviour of DR participants and a deeper comprehension of the factors driving response behaviour can assist in alleviating these challenges with better targeting and selection strategies of consumers for DR and the design of better-tailored schemes.

In this work, the authors approach the study of modelling and analysis of response behaviour under DR by utilising not primary load data (which is mainly used in such studies), but datasets related to household characteristics, when responding in an incentive-based DR programme. To this end, the study makes use of data from one of the largest open-source DR trial projects involving residential (household) consumers. We explored the datasets and pre-processed them accordingly (i.e. check for multi-collinearity, handling of missing values, outlier/anomaly detection and filtering). We explored the possibility of participating households learning to respond better over time, but we did not find supporting evidence for this hypothesis (at least for this specific DR scheme and available dataset).

We have showed that there is a statistically significant relationship between a household's response and a number of households' characteristics, and we have applied various machine learning techniques in search of better prediction accuracy. The prediction accuracy of these models found by our study is in line with expectations, given previous work on prediction of response behaviour which only focused on load data. Future work could combine both types of data this (household-related features and load data), potentially yielding better prediction accuracy for residential response behaviour.

We further explored and found the important features for modelling the households' response behaviour using two model-agnostic methods, across the different machine learning models. Next, we examined the effects of these important features on the households' response behaviour by using statistical distribution plots and interpretable AI methods based on coalitional game theory. We found that households with internet access, air-conditioning systems, power-intensive appliances (e.g. clothes dryer), and with lower gas usage tend to respond better than average. Moreover, there are also some insights which raise questions about the reported level of consumers' engagement in DR schemes, and the individual rationale of customers' response to DR signals.

Finally, taking a longer-term view, we argue that data-driven studies such as ours can provide key insights into better design of future residential DR trials, such that the data they collect can fill gaps in the knowledge of researchers and practitioners. For example, in this study, we could not detect a temporal trend showing an improvement in the response of households, as they participate in more DR events over time. This is likely due to the limited number of DR events the dataset from the SGSC trial (after filtering, we found only 10 events had reliable, high-quality, usable data). Another question that cannot be answered based on this data is what is the elasticity of the expected percentage demand reduction (response), w.r.t. the size of the financial reward of-

fered by the DR aggregator or network operator. In practice we expect this question to be important in designing, budgeting and optimising and budgeting residential DR schemes in the future.

Overall, with the roll-out of ubiquitous smart meters and large-scale availability of consumer-level response data, we argue data-driven methods will have an increasingly key role to play in the design of successful and fair DR schemes, with high participant engagement.

Data availability

The Smart-Grid Smart-City Customer Trial Data used in this study are available at the Australian Government data website: https://data.gov.au/dataset/ds-dga-4e21dea3-9b87-4610-94c7-15a8a77907ef/details.

Code availability

The reproducible code for this analysis can be found at: https://github.com/antongiannis/dr-behaviour-modelling-residential.git.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to acknowledge the support of the Energy Technology Partnership Scotland (ETP) through their Industry Doctorates scheme and our industrial sponsor Upside Energy. The work was also supported by the UK Engineering and Physical Sciences Council (EPSRC) through the UK National Centre for Energy Systems Integration (CESI) [EP/P001173/1] Community Energy Demand Reduction in India (CEDRI) [EP/R008655/1] and by Innovate UK through the Responsive Flexibility (ReFlex) project [ref:104780].

Appendix A

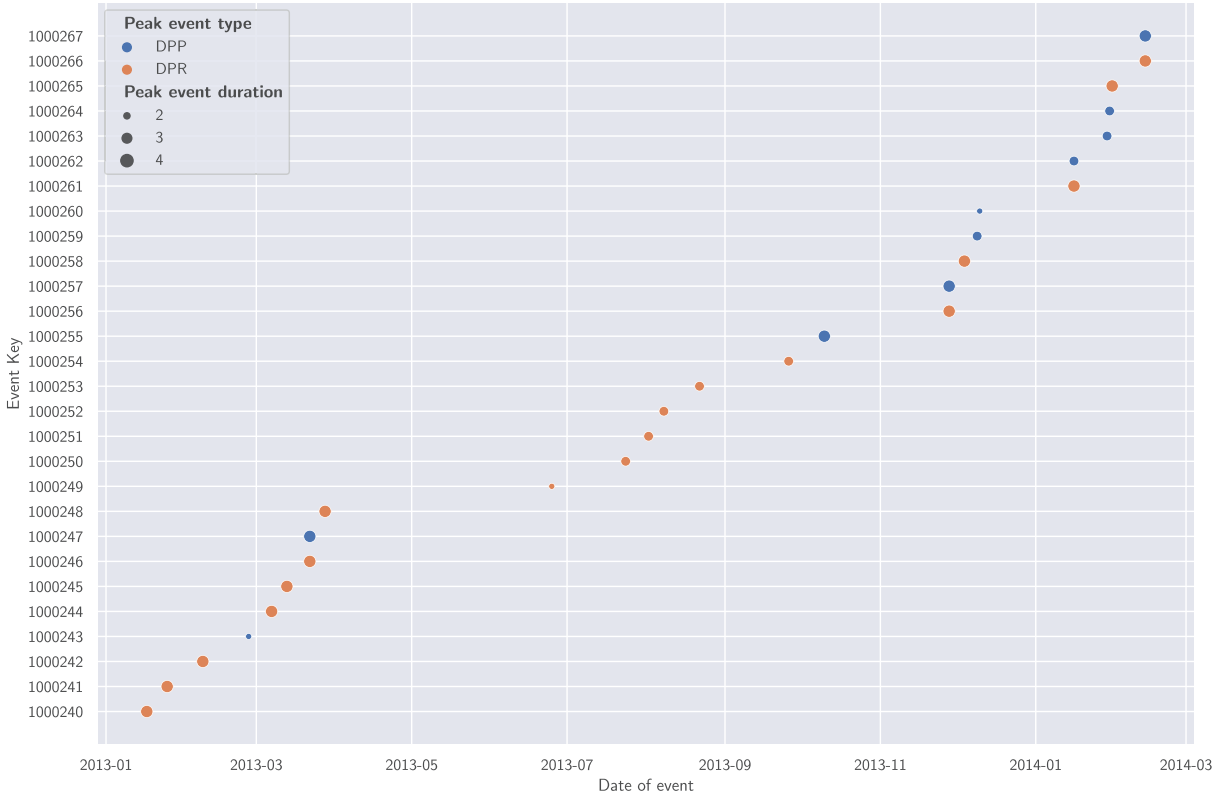


Fig. 9. Timestamp and duration of peak events of SGSC trials.

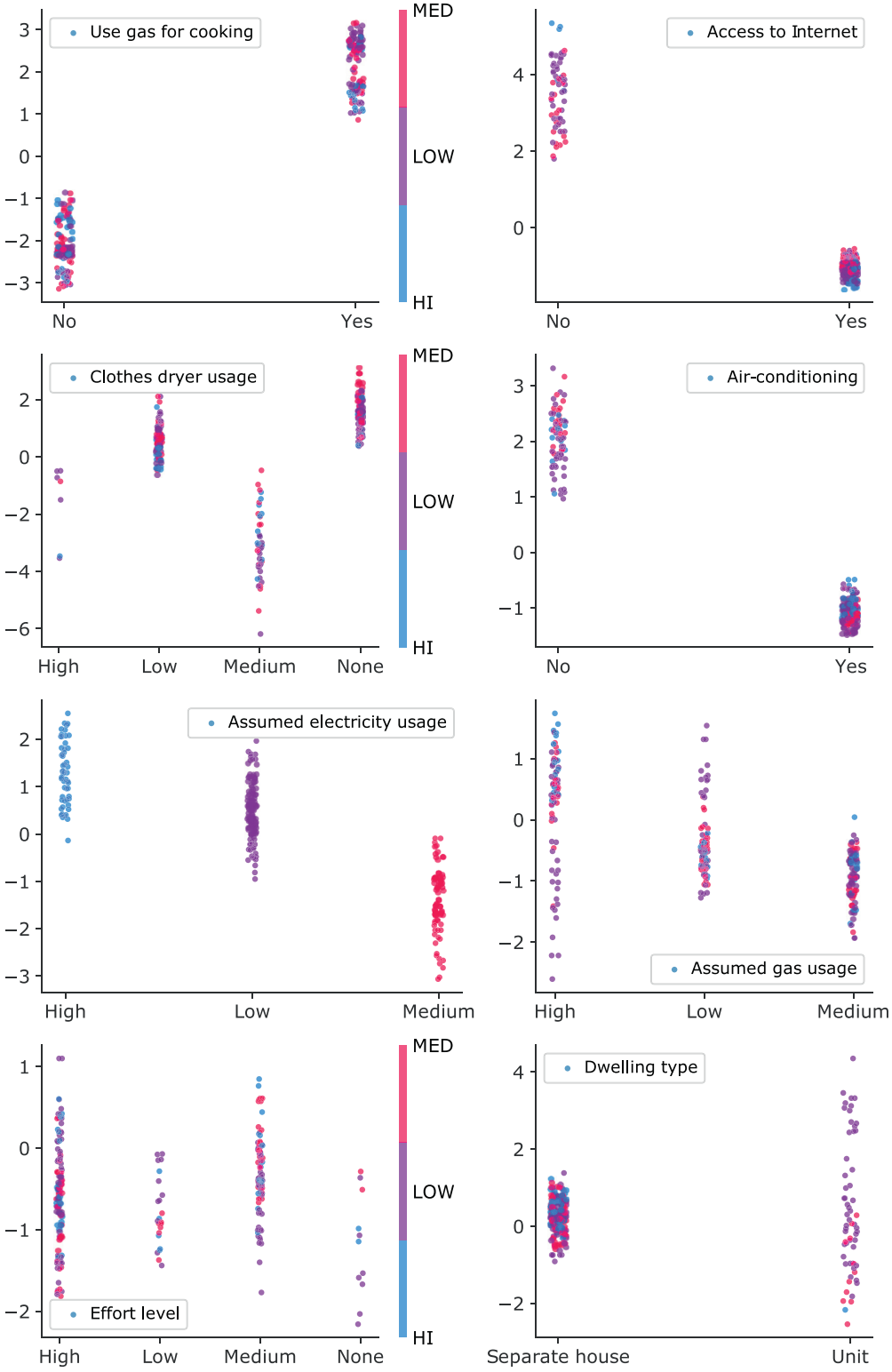


Fig. 10. SHAP values for important categorical features in conjunction with the electricity usage level. The colorbars indicate the electricity usage levels (HI, MED, LOW) of high, medium, and low.

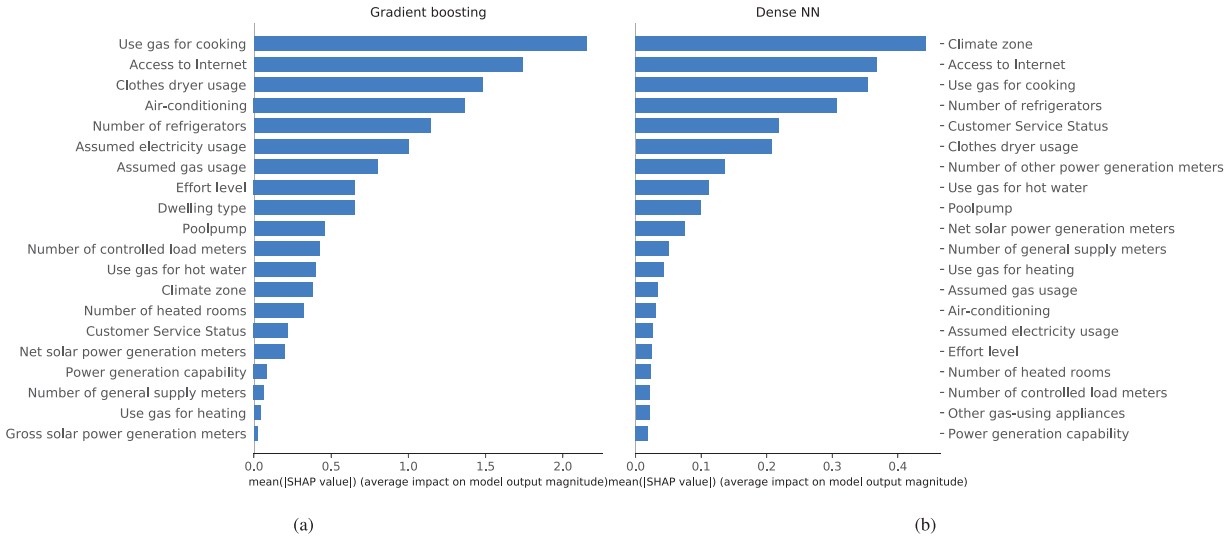


Fig. 11. Mean SHAP values of household characteristics based on Gradient boosting (a) and dense NN (b).

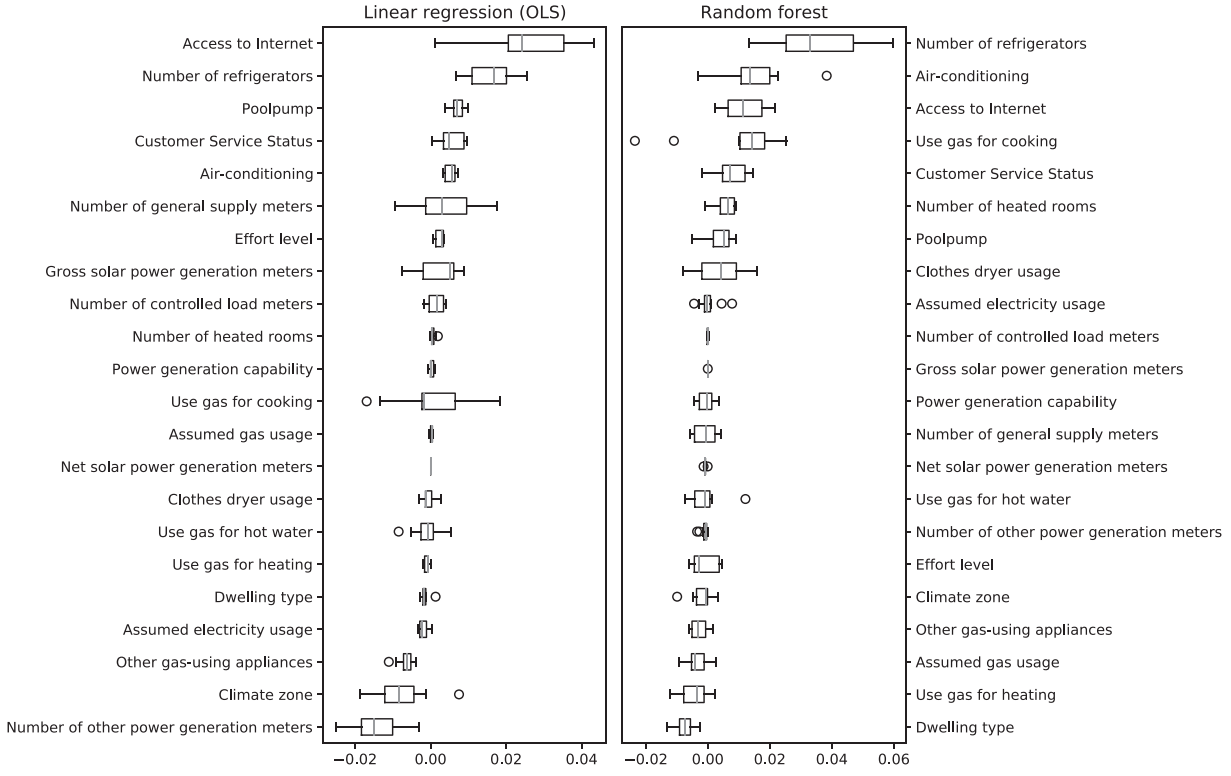


Fig. 12. Permutation feature importance results for the OLS regression and Random forest.

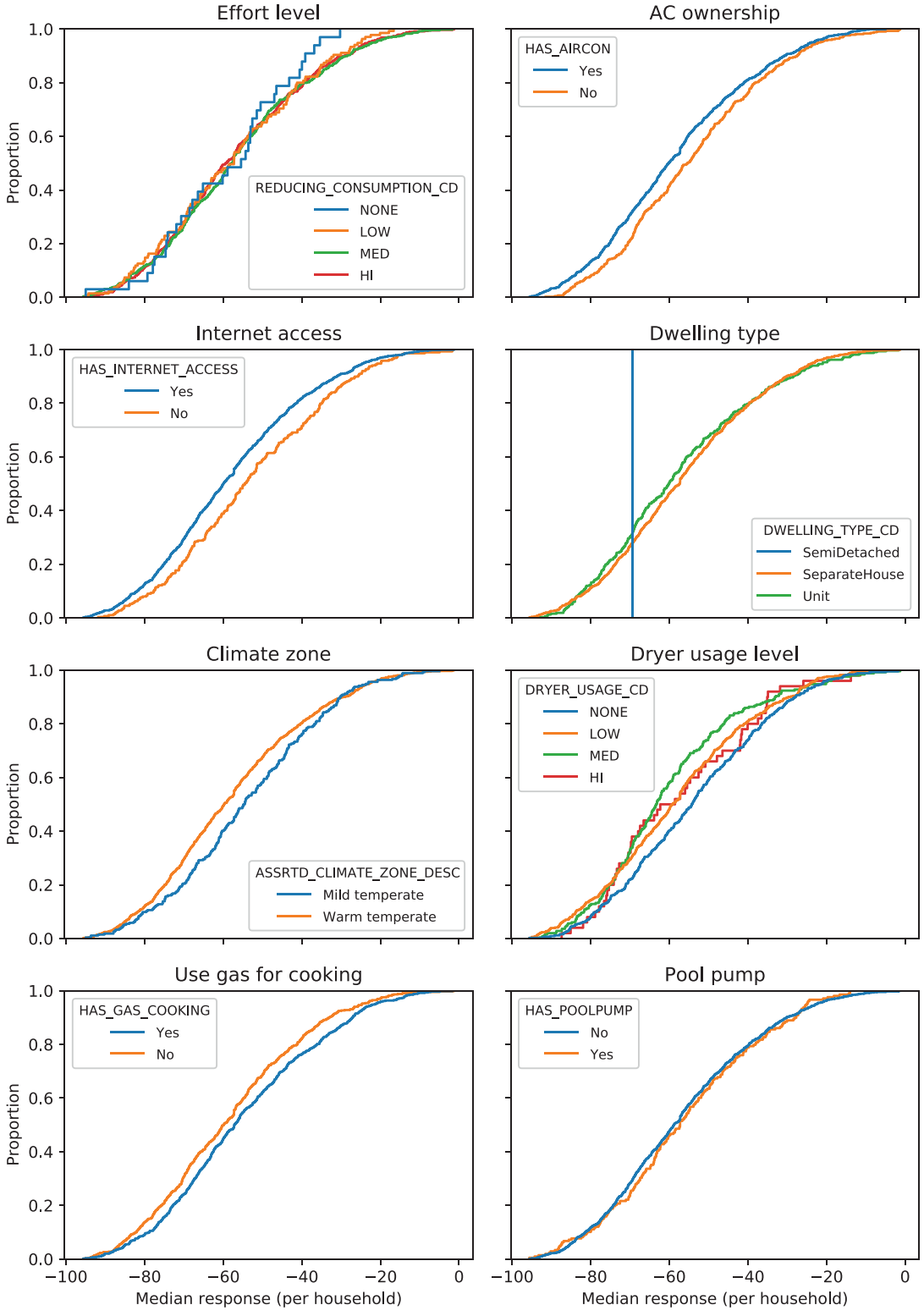


Fig. 13. Empirical CDF for the most important features.

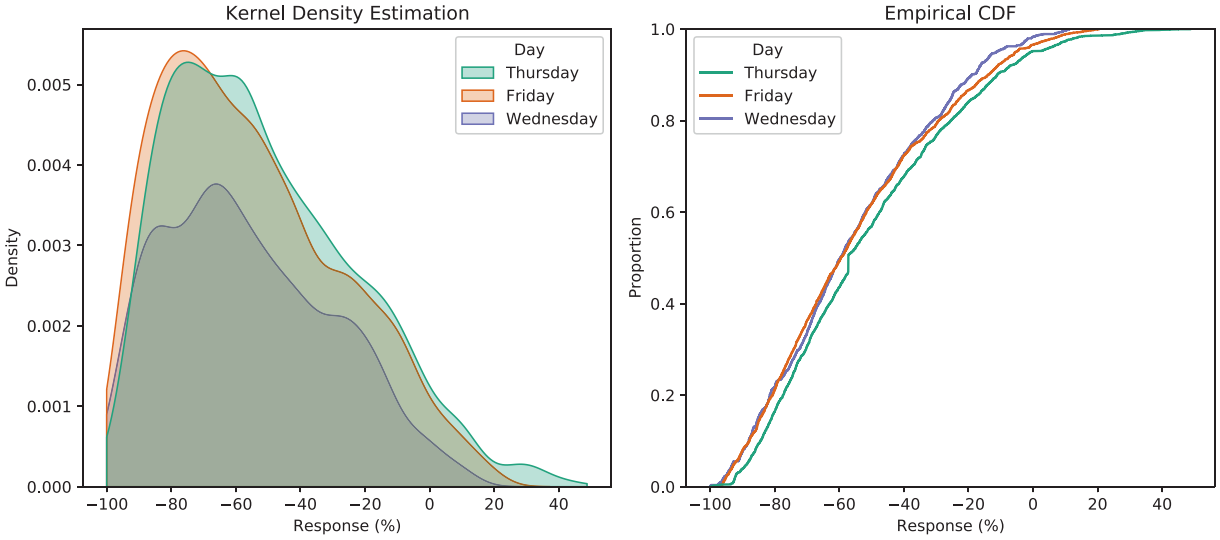


Fig. 14. Distribution plots of DPR responses for distinct days.

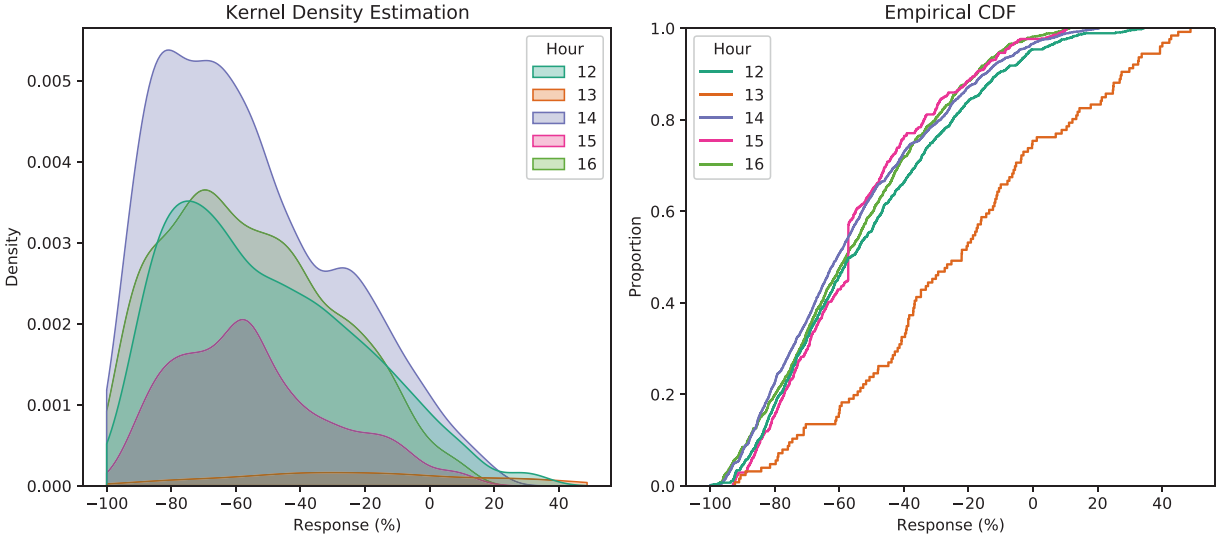


Fig. 15. Distribution plots of DPR responses for distinct hours of the day.

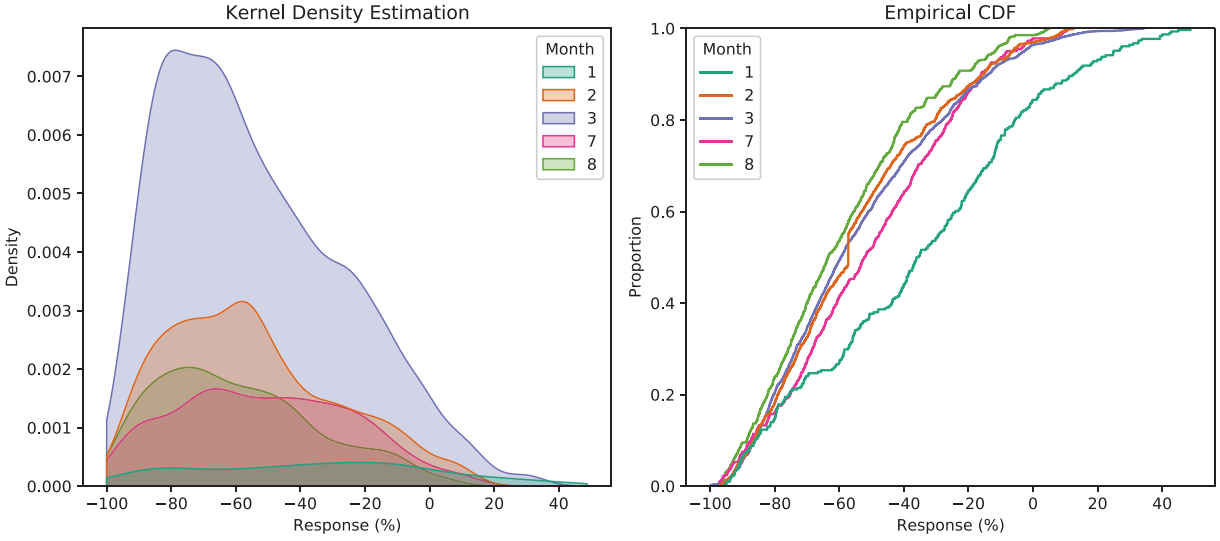


Fig. 16. Distribution plots of DPR responses for distinct months of the year.

References

[1] Eid C, Codani P, Perez Y, Reneses J, Hakvoort R. Managing electric flexibility from Distributed Energy Resources: a review of incentives for market design. Renew Sustain Energy Rev 2016;64:237–47. doi:10.1016/j.rser.2016.06.008. http://www.sciencedirect.com/science/article/pii/S1364032116302222

[2] Luo X, Wang J, Dooner M, Clarke J. Overview of current development in electrical energy storage technologies and the application potential in power system operation. Appl Energy 2015;137:511–36. doi:10.1016/J.APENERGY.2014.09.081. https://www.sciencedirect.com/science/article/pii/S0306261914010290

[3] IEA. Demand response (More efforts needed). Tech. Rep.. Paris: IEA; 2020. https://www.iea.org/reports/demand-response

[4] Faruqui A, Sergici S. Household response to dynamic pricing of electricity: a survey of 15 experiments. J Regul Econ 2010;38(2):193–225. doi:10.1007/s11149-010-9127-y.

[5] Jessoe K, Rapson D. Knowledge is (Less) power: experimental evidence from residential energy use. Am Econ Rev 2014;104(4):1417–38. doi:10.1257/aer.104.4.1417.

[6] Newsham GR, Bowker BG. The effect of utility time-varying pricing and load control strategies on residential summer peak electricity use: a review. Energy Policy 2010;38(7):3289–96. doi:10.1016/j.enpol.2010.01.027. https://www.sciencedirect.com/science/article/pii/S0301421510000510

[7] Herter K. Residential implementation of critical-peak pricing of electricity. Energy Policy 2007;35(4):2121–30. doi:10.1016/j.enpol.2006.06.019. https://www.sciencedirect.com/science/article/pii/S0301421506002783

[8] Liu D, Sun Y, Qu Y, Li B, Xu Y. Analysis and accurate prediction of user's response behavior in incentive-based demand response. IEEE Access 2019;7:3170–80. doi:10.1109/ACCESS.2018.2889500. https://ieeexplore.ieee.org/document/8587167/

[9] Ma Z, Xie J, Li H, Sun Q, Si Z, Zhang J, et al. The role of data analysis in the development of intelligent energy networks. IEEE Netw 2017;31(5):88–95. doi:10.1109/MNET.2017.1600319.

[10] Department of the Environment and Energy. Smart-grid smart-city customer trial data. 2015. https://data.gov.au/dataset/ds-dga-4e21dea3-9b87-4610-94c7-15a8a77907ef/details.

[11] Langham E, Downes J, Brennan T, Fyfe J, Mohr S, Rickwood P, White S. Smart grid, smart city customer research report. Tech. Rep.. Sydney: Report prepared by the Institute for Sustainable Futures as part of the AEFI consortium for Ausgrid and EnergyAustralia; 2014.

[12] Motlagh O, Foliente G, Grozev G. Knowledge-mining the Australian smart grid smart city data: a statistical-neural approach to demand-response analysis. Cham: Springer International Publishing; 2015. p. 189–207. ISBN 978-3-319-18368-8

[13] Kotu V, Deshpande B. Chapter 3 – data exploration. In: Kotu V, Deshpande B, editors. Predictive analytics and data mining. Boston: Morgan Kaufmann; 2015. p. 37–61. doi:10.1016/B978-0-12-801460-8.00003-3. ISBN 978-0-12-801460-8

[14] Schafer JL. Analysis of incomplete multivariate data. 1st ed. New York: CRC press; 1997. doi:10.1201/9780367803025. ISBN 9780367803025

[15] White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. Stat Med 2010;29(28):2920–31. doi:10.1002/sim.3944.

[16] Zhang S, Zhang C, Yang Q. Data preparation for data mining. Appl Artif Intell 2003;17(5–6):375–81. doi:10.1080/713827180.

[17] Kotu V, Deshpande B. Chapter 2 – data mining process. In: Kotu V, Deshpande B, editors. Predictive analytics and data mining. Boston: Morgan Kaufmann; 2015. p. 17–36. doi:10.1016/B978-0-12-801460-8.00002-1. ISBN 978-0-12-801460-8

[18] Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat 2001;29(5):1189–232. http://www.jstor.org/stable/2699986

[19] Hastie T, Tibshirani R, Friedman JHJH. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. Springer-Verlag New York; 2009. doi:10.1007/978-0-387-84858-7. ISBN 9780387848587

[20] Breiman L. Random forests. Machine Learn 2001;45(1):5–32. doi:10.1023/A:1010933404324.

[21] Russel S, Norvig P. Artificial intelligence a modern approach. 3rd ed; 2012. doi:10.1017/S0269888900007724. ISBN 0136042597, eprint 9809069v1

[22] Goodfellow I, Bengio Y, Courville A. Deep learning. MIT Press; 2016.

[23] Kim B, Khanna R, Koyejo OO. Examples are not enough, learn to criticize! Criticism for Interpretability. In: Lee D, Sugiyama M, Luxburg U, Guyon I, Garnett R, editors. Advances in neural information processing systems, 29. Curran Associates, Inc.; 2016. p. 2280–8. https://proceedings.neurips.cc/paper/2016/file/5680522b8e2bb01943234bce7bf84534-Paper.pdf

[24] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, editors. Advances in neural information processing systems 30. Curran Associates, Inc.; 2017. p. 4765–74. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[25] Altmann A, Toloşi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. Bioinformatics 2010;26(10):1340–7. doi:10.1093/bioinformatics/btq134.

[26] Strobl C, Boulesteix A-L, Zeileis A, Hothorn T. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinform 2007;8(1):25. doi:10.1186/1471-2105-8-25.

[27] Chalkiadakis G, Elkind E, Wooldridge M. Computational aspects of cooperative game theory. Synth Lect Artif IntellMach Learn 2011;5(6):1–168. doi:10.2200/S00355ED1V01Y201107AIM016.

[28] Antonopoulos I, Robu V, Couraud B, Kirli D, Norbu S, Kiprakis A, et al. Artificial intelligence and machine learning approaches to energy demand-side response: A systematic review. Renew Sustain Energy Rev 2020;130(109899):35. doi:10.1016/j.rser.2020.109899.

[29] Zhou D, Balandat M, Tomlin C. A Bayesian perspective on Residential Demand Response using smart meter data. In: 2016 54th annual allerton conference on communication, control, and computing (Allerton). IEEE; 2016. p. 1212–19. doi:10.1109/ALLERTON.2016.7852373. ISBN 978-1-5090-4550-1

[30] Paterakis NG, Tascikaraoglu A, Erdinc O, Bakirtzis AG, Catalao JPS. Assessment of demand-response-driven load pattern elasticity using a combined approach for smart households. IEEE Trans Ind Inform 2016;12(4):1529–39. doi:10.1109/TII.2016.2585122. http://ieeexplore.ieee.org/document/7500042/

[31] Holtschneider T, Erlich I. Optimization of electricity pricing considering neural network based model of consumers' demand response. In: 2013 IEEE Computational Intelligence Applications in Smart Grid (CIASG). IEEE; 2013. p. 154–60. doi:10.1109/CIASG.2013.6611512. ISBN 978-1-4673-6002-9

[32] Kang J, Lee S. Data-driven prediction of load curtailment in incentive-based demand response system. Energies 2018;11(2905):11. doi:10.3390/en11112905.

[33] Cai L, Wen H, Gu J, Ma J, Jin Z. Forecasting customers' response to incentives during peak periods: a transfer learning approach. Int Trans Electr Energy Syst 2020;30(7):e12251. doi:10.1002/2050-7038.12251.

[34] Varghese AC, V P, Kumar G, Khaparde SA. Smart grid consumer behavioral model using machine learning. In: 2018 IEEE innovative smart grid technologies – Asia (ISGT Asia). IEEE; 2018. p. 734–9. doi:10.1109/ISGT-Asia.2018.8467824. ISBN 978-1-5386-4291-7

[35] Haben S, Singleton C, Grindrod P. Analysis and clustering of residential customers energy behavioral demand using smart meter data. IEEE Trans Smart Grid 2016;7(1):136–44. doi:10.1109/TSG.2015.2409786. http://ieeexplore.ieee.org/document/7063233/

[36] Aïd R., Possamaï D., Touzi N.. Optimal electricity demand response contracting with responsiveness incentives. 2019. 1810.09063.

[37] Élie R, Hubert E, Mastrolia T, Possamaï D. Mean field moral hazard for optimal energy demand response management. Math Financ 2021;31(1):399–473. doi:10.1111/mafi.12291. https://onlinelibrary.wiley.com/doi/abs/10.1111/mafi.12291

[38] Meir R, Ma H, Robu V. Contract design for energy demand response. In: Proceedings of the twenty-sixth international joint conference on artificial intelligence, IJCAI-17; 2017. p. 1202–8. doi:10.24963/ijcai.2017/167.

[39] Ma H, Robu V, Li N, Parkes DC. Incentivizing reliability in demand-side response. In: IJCAI 2016: 25th international joint conference on artificial intelligence; 2016. p. 352–8.

[40] Ma H, Parkes DC, Robu V. Generalizing demand response through reward bidding. In: Proceedings of the 16th international confer- ence on autonomous agents and multiagent systems.. ACM; 2017. p. 60–8. ISBN 9781510855076

[41] Kota R, Chalkiadakis G, Robu V, Rogers A, Jennings NR. Cooperatives for demand side management. In: Proceedings of the 20th European conference on artificial intelligence; 2012. p. 969–74. doi:10.3233/978-1-61499-098-7-969. https://dl.acm.org/citation.cfm?id=3007521

[42] Norbu S, Couraud B, Robu V, Andoni M, Flynn D. Modelling the redistribution of benefits from joint investments in community energy projects. Appl Energy 2021;287:116575. doi:10.1016/j.apenergy.2021.116575. https://www.sciencedirect.com/science/article/pii/S0306261921001215

[43] Robu V, Vinyals M, Rogers A, Jennings NR. Efficient buyer groups with prediction-of-use electricity tariffs. IEEE Trans Smart Grid 2018;9(5):4468–79. doi:10.1109/TSG.2017.2660580. https://ieeexplore.ieee.org/document/7835716/

[44] Robu V, Gerding EH, Stein S, Parkes DC, Rogers A, Jennings NR. An online mechanism for multi-unit demand and its application to plug-in hybrid electric vehicle charging. J Artif Intell Res 2013;48:175–230. doi:10.1613/JAIR.4064. https://www.jair.org/index.php/jair/article/view/10838

[45] Stein S, Gerding E, Robu V, Jennings NR. A model-based online mechanism with pre-commitment and its application to electric vehicle charging. In: Proc. of the 11th int. conference on autonomous agents and multiagent systems - Volume 2. AAMAS '12; 2012. p. 669–76. ISBN 0981738125

[46] Cramér H. The two-dimensional case. Princeton University Press; 1946. ISBN 0-691-08004-6

[47] Vatcheva KP, Lee M, McCormick JB, Rahbar MH. Multicollinearity in regression analyses conducted in epidemiologic studies. Epidemiology 2016;6(2):227. doi:10.4172/2161-1165.1000227. https://pubmed.ncbi.nlm.nih.gov/27274911 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4888898/

[48] Basu I., Maji S.. Multicollinearity correction and combined feature effect in shapley values. 2020. 2011.01661.

[49] Nolan S, O'Malley M. Challenges and barriers to demand response deployment and evaluation. Appl Energy 2015;152:1–10. doi:10.1016/j.apenergy.2015.04.083. http://www.sciencedirect.com/science/article/pii/S0306261915005462

[50] Weng Y, Rajagopal R. Probabilistic baseline estimation via Gaussian process. In: 2015 IEEE Power & Energy Society General Meeting. IEEE; 2015. p. 1–5. doi:10.1109/PESGM.2015.7285756. ISBN 978-1-4673-8040-9

[51] Breunig MM, Kriegel H-P, Ng RT, Sander J. LOF: identifying density-based local outliers. SIGMOD Rec 2000;29(2):93–104. doi:10.1145/335191.335388.

[52] Liu FT, Ting KM, Zhou Z-H. Isolation-based anomaly detection. ACM Trans Knowl Discov Data 2012;6(1). doi:10.1145/2133360.2133363.

[53] Cheng Z, Zou C, Dong J. Outlier detection using isolation forest and local outlier factor. In: Proceedings of the conference on research in adaptive and convergent systems. New York, NY, USA: Association for Computing Machinery; 2019. p. 161–8. doi:10.1145/3338840.3355641. ISBN 9781450368438

[54] Ljung GM, Box GEP. On a measure of lack of fit in time series models. Biometrika 1978;65(2):297–303. doi:10.2307/2335207. http://www.jstor.org/stable/2335207

[55] Escanciano JC, Lobato IN. An automatic Portmanteau test for serial correlation. J Econ 2009;151(2):140–9. doi:10.1016/j.jeconom.2009.03.001. http://www.sciencedirect.com/science/article/pii/S0304407609000773
[56] Murphy KP. Machine learning: a probabilistic perspective. MIT Press; 2012. ISBN 9780262018029
[57] Seabold S, Perktold J. Statsmodels: econometric and statistical modeling with python. In: 9th Python in science conference; 2010. p. 92–6.
[58] Prokhorenkova L., Gusev G., Vorobev A., Dorogush A.V., Gulin A.. CatBoost: unbiased boosting with categorical features. 2019. 1706.09516.
[59] Micci-Barreca D. A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. SIGKDD Explor Newsl 2001;3(1):27–32. doi:10.1145/507533.507538.
[60] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825–30.
[61] Abadi M., Agarwal A., Barham P., Brevdo E., Chen Z., Citro C., Corrado G.S., Davis A., Dean J., Devin M., Ghemawat S., Goodfellow I., Harp A., Irving G., Isard M., Yangqing J., Jozefowicz R., Kaiser L., Kudlur M., Levenberg J., Mané D., Monga R., Moore S., Murray D., Olah C., Schuster M., Shlens J., Steiner B., Sutskever I., Talwar K., Tucker P., Vanhoucke V., Vasudevan V., Viégas F., Vinyals O., Warden P., Wattenberg M., Wicke M., Yu Y., Zheng X.. TensorFlow: large-scale machine learning on heterogeneous systems. 2015. https://www.tensorflow.org/.
[62] Kingma DP, Ba J. Adam: a method for stochastic optimization. In: Bengio Y, LeCun Y, editors. 3rd international conference on learning representations, {ICLR} 2015, San Diego, CA, USA, May 7-9, 2015, conference track proceedings; 2015. http://arxiv.org/abs/1412.6980
[63] O'Malley T., Bursztein E., Long J., Chollet F., Jin H., Invernizzi L., Others. Keras Tuner. 2019.
[64] Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Mach Learn Res 2012;13(null):281–305.
[65] Hancock JT, Khoshgoftaar TM. Survey on categorical data for neural networks. J Big Data 2020;7(1):28. doi:10.1186/s40537-020-00305-w.
[66] Dwork C, Kumar R, Naor M, Sivakumar D. Rank aggregation methods for the web. In: Proceedings of the 10th international conference on world wide web. New York, NY, USA: Association for Computing Machinery; 2001. p. 613–22. doi:10.1145/371920.372165. ISBN 1581133480