Programme Area: Smart Systems and Heat

Project: WP1 Appliance Disaggregation

Title: High Frequency Appliance Disaggregation Analysis: Insights Overview

Abstract:

The ETI collected utility meter and other data (e.g. room temperatures, humidity, and HEMS control data) from five dwellings over a period of six months. Using the collected data, work was conducted to evaluate different machine learning algorithms, research appropriate data features and calibrations thereof, and test the 'art of the possible'. This report was commissioned in order to share the key learnings from the ETI High Frequency Appliance Disaggregation Analysis work carried out jointly by Baringa Partners and ASI Data Science. At a high level, the work evaluates machine learning algorithms, researches potential data features and calibrations thereof, and tests the "art of the possible" when it comes to predicting future occupancy and hot water usage from multi-vector high frequency meter data. Crucially, the work conducted not only seeks to understand historical human activity within the residency, but also to predict future needs. This report serves the purpose of sharing the key insights of the work to date as well as sharing areas of potential future work.

Context:

The High Frequency Appliance Disaggregation Analysis (HFADA) project builds upon work undertaken in the Smart Systems and Heat (SSH) programme delivered by the Energy Systems Catapult for the ETI, to refine intelligence and gain detailed smart home energy data. The project analysed in depth data from five homes that trialed the SSH programme's Home Energy Management System (HEMS) to identify which appliances are present within a building and when they are in operation. The main goal of the HFADA project was to detect human behaviour patterns in order to forecast the home energy needs of people in the future. In particular the project delivered a detailed set of data mining algorithms to help identify patterns of building occupancy and energy use within domestic homes from water, gas and electricity data.

Disclaimer: The Energy Technologies Institute is making this document available to use under the Energy Technologies Institute Open Licence for Materials. Please refer to the Energy Technologies Institute website for the terms and conditions of this licence. The Information is licensed 'as is' and the Energy Technologies Institute excludes all representations, warranties, obligations and liabilities in relation to the Information to the maximum extent permitted by law. The Energy Technologies Institute is not liable for any errors or omissions in the Information and shall not be liable for any loss, injury or damage of any kind caused by its use. This exclusion of liability includes, but is not limited to, any direct, indirect, special, incidental, consequential, punitive, or exemplary damages in each case such as loss of revenue, data, anticipated profits, and lost business. The Energy Technologies Institute does not guarantee the continued supply of the Information. Notwithstanding any statement to the contrary contained on the face of this document, the Energy Technologies Institute confirms that it has the right to publish this document.

High Frequency Appliance Disaggregation Analysis: Insights Overview

CLIENT: Energy Technologies Institute

DATE: 14/05/2018

Baringa Partners LLP is a Limited Liability Partnership registered in England and Wales with registration number OC303471 and with registered offices at 3rd Floor, Dominican Court, 17 Hatfields, London SE1 8DJ UK.

Version History

Version	Date	Description	Prepared by	Approved by
1.0	14/05/2018	Final Insights Overview document	Alberto Favaro, Zhihan Xu, Cris Lowery	Oliver Rix

Contact

Name Oliver.Rix@baringa.com +44 7790 017 576

Confidentiality and Limitation Statement

This document is provided to the ETI under, and is subject to the terms of, the Energy Technologies Institute’s Agreement for the High Frequency Appliance Disaggregation Project.

Contents

Figures

Tables

1 Executive Summary

This report was commissioned in order to share the key learnings from the ETI High Frequency Appliance Disaggregation Analysis work carried out jointly by Baringa Partners and ASI Data Science. At a high level, the work evaluates machine learning algorithms, researches potential data features and calibrations thereof, and tests the "art of the possible" when it comes to predicting future occupancy and hot water usage from multi-vector high frequency meter data. Crucially, the work conducted not only seeks to understand historical human activity within the residency, but also to predict future needs. This report serves the purpose of sharing the key insights of the work to date as well as sharing areas of potential future work.

Given the project was a research project the scope was not fixed, but instead mechanisms to regularly review the approach and scope with ETI were employed, including a fortnightly project management and sprint call. The team largely drew on an agile delivery methodology to ensure that learnings from the previous sprints were easily leveraged in future sprints. Whilst it was important to quickly adapt the scope, it was also important to provide a broad direction for the project activities. For this reason, the project was broken down at the start into five key sub-stages, which are: data engineering, data exploration, evolution and iteration, integration, and documentation.

The results of the work are promising suggesting that both occupancy and hot water usage are predictable within at least a 24-hour time horizon, but the results need further validation given that the predictive models have so far only been tested on one property for one week. Table 1 provides an overview of the predictive performance results as measured by the Area Under the Curve (AUC), which highlights that the 4 hour performance is 78% for occupancy and 62% for hot water usage. The predictive algorithm used in the work is a Random Forest Classifier due to its robustness, high predictive performance and the ability to interpret feature importance. The final features selected and the calibration of the model was achieved through k-fold cross validation. It is interesting to note that the occupancy model outperforms the hot water usage one throughout, which is likely due to occupancy being a binary rather than trinary problem and that it is arguably more stable and consistent over the timeframes analysed.

Time horizon	Occupancy performance [AUC]	Hot water usage performance [AUC]
10 mins	98%	92%
1 hr	89%	70%
4 hrs	78%	62%
24 hrs	63%	61%
72 hrs	69%	46%

Table 1: Predictive performance of various models

The report also elaborates on the most predictive data features for each of the models. For both the hot water usage models and the occupancy models, using the historical and current value of the target variable proved to be valuable, especially at the shorter time horizons. It was also found that bathroom humidity and electricity principal components had high predictive power for both model sets. Interestingly, the 25 electricity state clusters generated from the electricity principal components are not found to add much predictive performance on top of the 50 electricity principal components, which is probably due to the Random Forest being able to capture the relevant structure directly from the principal components. However, the clusters were useful in generating the occupancy label and future work may find the representation useful in analysing resident workflow and capturing memory in a more powerful or efficient way. Finally, memory is found to be key and was introduced in various data features. The time span of memory is found to be powerful when comparable to the time horizon over which one is predicting or at 24 hours. The key data feature that was powerful for the hot water usage model and not for the occupancy model is water usage. In contrast, exogenous factors seemed to be much more powerful for the occupancy model than the hot water usage one.

Given that the predictive models have only been tested on 1 week of data for 1 property, it is important that future work focuses on extending the analysis to a larger dataset. It is also suggested, that further data feature engineering could improve performance, for example capturing human workflows or running appliance disaggregation analysis. Suggestions are also made in terms of the choice of predictive algorithm and how the problem is framed. Whilst out of the scope of this work, it may be interesting to consider extensions that focus on how to reduce the computational time or cost involved in obtaining the predictive performance, how to automate the human intervention in a production environment, and evaluate the commercial value of the solution.

2 Introduction

At the time of writing, the ETI is investigating the development of a Home Energy Management System (HEMS) capable of optimising the comfort of a dwelling's residents while managing the necessary energy expenditure. As part of this initiative, it is investigating a system that can learn future patterns of occupancy and needs of its residents using non-intrusive monitoring equipment from two or more utilities. Three key differentiators of the work undertaken, as compared to prior "Non-Intrusive Appliance Load Monitoring" research, are:

1. Monitoring multiple utilities to provide more information and contextual knowledge;
2. Recording high frequency electricity data to provide additional information on current property state;
3. Potential use of priors to more effectively identify behavioural patterns of property states.

To facilitate the research, the ETI collected utility meter consumption data and other data (e.g. humidity, and HEMS control data) from five dwellings over a period of six months. Using a subset of the collected data, work has been conducted to evaluate machine learning algorithms, research potential data features and calibrations thereof, and test the "art of the possible".

This report explains the high level methodology followed, and the limitations thereof, providing detail on some of the key design choices including the more complex data features, summarising the results achieved for the predictive models and describing areas of potential future work.

To start with, Chapter 4 explains some of the key methodological choices and the limitations thereof. Initially, there is a strong focus on the data used, the limitations of the subset of data employed and the pre-processing that was required. The section then describes the framework used for optimising and testing the model, including balancing classes, measuring performance and splitting the data into a training, validation and test set. Finally, the choice of algorithm and the framing of the predictive model is explained.

Another key area of focus of the report is the feature engineering. In Chapter 5, the need for data compression of the electricity data is explained as well as the approach to data compression and the reasoning for this. The chapter also describes the key data features that were developed in order to enhance predictive performance paying particular attention to electricity clustering, memory, exogenous factors and auto-regressive terms.

Following the detailing of the methodology and the feature engineering, the results are presented for the hot water usage models and the occupancy models in Chapters 6 and 7, respectively. The results are broken down by the five time horizons and focus on the performance of the model, as measured by the Area Under the Curve (AUC), and the most predictive data features.

Finally, the report concludes with ideas of potential future work in order to improve the current analysis and the predictive power of the results. Chapter 8 recommends four broad areas of focus: enhancing the dataset, generating more data features, testing alternative machine learning algorithms and framing the predictive problem in a different way.

3 Methodology

3.1 Data used

3.1.1 Data provided

Data was collected for five different properties for a period of approximately six months. At the stage of writing, the team have received 30 hard drives, in three batches of 10. However, in the first stage of the project, only the first batch had been received. This batch of hard drives was analysed for data quality issues, which helped the team understand the overall data quality and direct the research efforts more effectively. The below table provides an overview of the first 10 hard drives received and analysed.

	H20	*H25*	*H45*	*H71*	*H73*
No. of hard drives	1	3	3	2	1
Electricity data timespan	31 May – 3 Jul	21 Mar – 20 Jul	31 Mar – 4 Jul	28 Apr – 3 Jul	5 Jun – 3 Jul
Electricity data quality	Solar panels present	Frequent, repeated long gaps	~15 days missing; well-defined gaps	~2 days missing; well-defined gaps	~5 days missing; well-defined gaps
Water data timespan and quality	31 May – 3 Jul	21 Mar – 20 Jul (2 water meters)	30 Mar – 4 Jul. Good data quality	28 Apr – 3 Jul. Good data quality	5 Jun – 3 Jul. Good data quality
HEMS database 1	NA	20 Mar – 8 May	20 Mar – 8 May	25 Apr – 8 May	NA
HEMS database 2	Unaudited	Unaudited	Unaudited	Unaudited	Unaudited
Home survey & floor plans	Available	Available	Available	Available	Available

Table 2: Overview of data received by property for the first 10 hard drives

3.1.2 Data quality

The data quality checks conducted were not exhaustive and they were limited to the first 10 hard drives, but they did highlight several key points. The main risks and complications identified for using the full dataset are:

- it will be hard to use 1 property due to the presence of a solar panel;
- there are significant gaps in the data, shortening the useful time period;
- the HEMS data is stored in an unstructured way, leading to extra mapping work;
- there is significant drift in the time series, making it hard to perfectly line up the data;
- there may be additional data quality issues not identified through the predominant statistical test that were run, which focussed in on: data completeness, stuck values and load profiles;
- there may be additional data quality issues in the other 20 hard drives that were not investigated.

Based on the data completeness and the data quality checks from the initial 10 hard drives the following view of priority for analysis was generated:

1. H45: Has most electricity and water data available, with no major issues
2. H71: Has second most electricity and water data available, with no major issue. Water and electricity consumption appear a bit low
3. H73: Does not appear to have HEMS data (may be present in second database)
4. H25: Deprioritised due to frequent and long gaps in electricity data
5. H20: Deprioritised as it has solar panels, making analysis very difficult.

Analysing H45 was prioritised and the rest of the report is based on the insights gained from this property using the 30 days with a green overview in the table below.

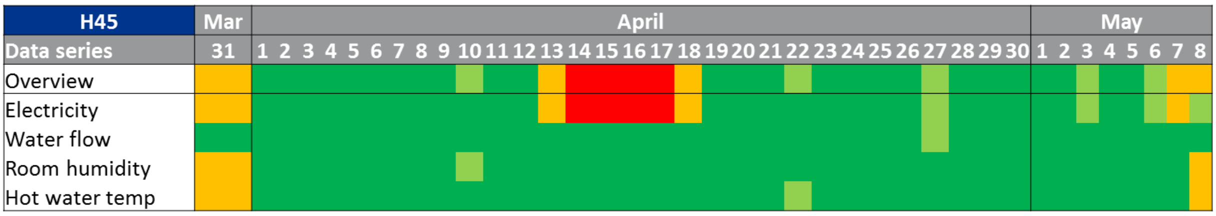

Table 3: Data completeness for H45 for first 39 days of data (9 days were lost due to electricity usage data gaps, 6 further days would have been lost if we included gas consumption data)

3.1.3 Limitations of experimental setup

The experimental setup, including the data used, has key limitations to it that are worth highlighting. These are:

- **Data sparseness**: the time range of data collected for this research is sensible as it balances both the need to have enough data to identify patterns and the need to test that behaviours can be learnt in a short enough period for a human to find such a solution useful. That said, it is important to acknowledge that certain human activity is not very repetitive such as holidays and buying new appliances, making it very hard to recognise these events. This issue has been worsened by the fact that the insights in this report were based on a subset of the original data collected (1 property for 30 days). It is important to consider how to handle these events in a future solution, which could include the use of priors.
- **Data quality issues**: some of these were highlighted in the previous section and include presence of a solar panel, missing data, noise in readings, time drift, etc. Such issues will affect the ability to analyse all the data as well as the precision and accuracy of the results obtained. Additionally, hot water usage is not explicitly captured, but inferred from pipe temperature and water usage leading to a source of inaccuracy.
- **Results validation**: a process was followed in which occupancy was inferred and labelled for each 10 minute period using a human interpretation of the electricity, water and pipe temperature activities observed in the property. This labelling exercise does not provide a ground truth, but an approximation thereof. As such, the reported predictive performance for

occupancy is an estimation of the true value. Given the human judgement involved, there may also be some biases introduced through the labelling process.

3.1.4 Pre-processing required

Given that data was required from multiple sources, recorded at different frequencies and stored in multiple file types, it was necessary to do some data pre-processing prior to any analysis. For full details please refer to the handover documentation. The steps involved included:

- investigating data quality and correcting for data quality issues;
- correcting for time drift;
- up-sampling and down-sampling to 1 second time intervals, as required;
- extracting power harmonics;
- applying PCA to harmonics;
- syncing across different data sources;
- clustering principal components;
- approximating hot water usage;
- inferring occupancy.

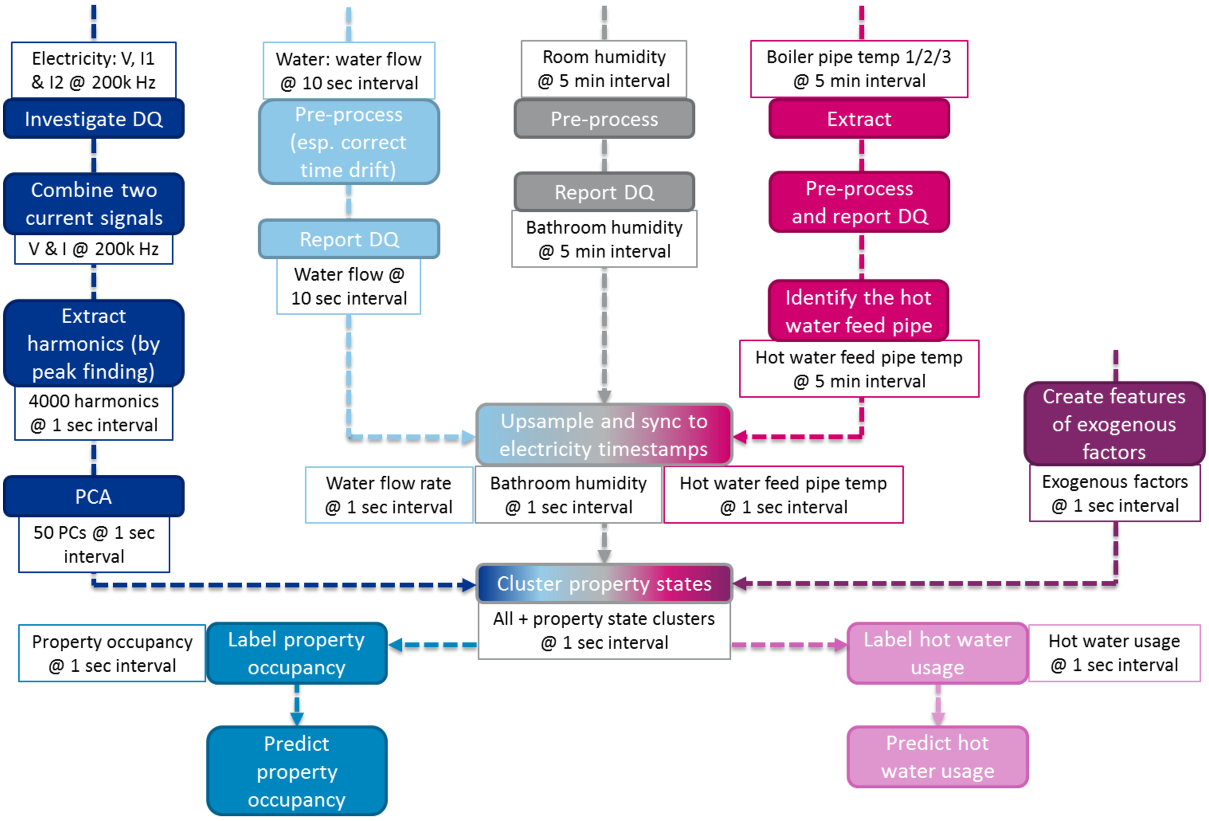
Electricity: V, I1 & I2 @ 200k Hz
Investigate DQ
Combine two current signals
V & I @ 200k Hz
Extract harmonics (by peak finding)
4000 harmonics @ 1 sec interval
PCA
50 PCs @ 1 sec interval
Water: water flow @ 10 sec interval
Pre-process (esp. correct time drift)
Report DQ
Water flow @ 10 sec interval
Room humidity @ 5 min interval
Pre-process
Report DQ
Bathroom humidity @ 5 min interval
Boiler pipe temp 1/2/3 @ 5 min interval
Extract
Pre-process and report DQ
Identify the hot water feed pipe
Hot water feed pipe temp @ 5 min interval
Upsample and sync to electricity timestamps
Water flow rate @ 1 sec interval
Bathroom humidity @ 1 sec interval
Hot water feed pipe temp @ 1 sec interval
Create features of exogenous factors
Exogenous factors @ 1 sec interval
Cluster property states
All + property state clusters @ 1 sec interval
Property occupancy @ 1 sec interval
Label property occupancy
Predict property occupancy
Label hot water usage
Hot water usage @ 1 sec interval
Predict hot water usage

Figure 1: Overview of data processing steps required to generate a unified property view

3.2 Framework for model optimisation and testing

3.2.1 Training, validation & testing datasets

When building a machine learning model it is important to use an appropriate framework for training, validating and testing to ensure that the model is appropriately optimised and that the right statistical performance is reported. Training data is required to train the model, the validation data is used to compare different parameterisations of the model and feature combinations, the test data allows testing of the model performance with new data providing an unbiased estimate of performance.

The results reported in this report are based on the following configuration:

- 30 days of data are used: 23 days for training & validating, 7 days for testing.
- The test set occurs after the training/validation one, as it is important to not train a model using future data to predict past results.
- The test set was chosen to be exactly 1 week, to avoid any within week biases.
- The hot water usage predictive model uses 3-fold cross-validation (purpose built script). The folds are not randomised but are a chunk of around 8 consecutive days.
- The occupancy predictive model uses 4-fold cross-validation using a standard implementation provided in the scikit-learn library.

3.2.2 Balancing classes

Some of the predictive models were highly imbalanced, meaning that one or more of the classes are underrepresented in the data. This makes it hard for predictive algorithms to predict these classes as there are less examples of these for the algorithm to train on. Therefore, it is common practice to closely balance the classes, which is also the approach taken on this project. In order to balance the classes, the various classes were up-sampled until a ratio of approximately 1:1 was reached in occupancy and 1:1:1 in hot water usage. The model was trained and validated on the balanced version of the data, but to give accurate performance results it was tested on a non up-sampled version of the test data.

3.2.3 Measuring performance

The validation datasets were used to measure performance of models using different hyper-parameters and features, in order to determine a good combination of these to produce a final model. A key challenge that is faced when measuring performance, is that there are two types of errors that need to be balanced in a binary predictive problem and several more for a multi-class problem. For instance, taking the binary classifier for ease, there is the possibility of predicting a positive case whilst the true value is negative, and predicting a negative case whilst the true value is positive. The relative impact of these two errors are dependent on customer preferences, and is known as a loss function. Given that our loss function will vary per user and is unknown at this stage, the work looked to pick the algorithm with the best overall range of performance, as measured by the Area Under the Curve (AUC) of a Receiver Operating Characteristic.

The algorithm used, Random Forest Classifier, also outputs a probability of each class. Using the probability associated to each class and the probability of each type of error, it is possible to incorporate a customer's loss function into the decisioning engine that is used to determine the right level of heating of the property and of the water. The methodology to do so is outside of the scope of this work, and would require further work in order to calibrate the output probabilities against true probabilities.

3.2.4 Volatility in performance

The data size used for training, validating and testing consisted of 30 days. Given that a significant amount of human behaviour will repeat at the day or week level and that in some cases forecasts were at the 72 hour level, this is quite a small amount of data for a high-dimensional machine learning algorithm. This places the machine learning algorithm at risk of over-fitting (fitting very well to the intricacies in the training/validation data, but failing to generalise well to new data) and also encountering new behaviours in the test set that were not observed in the training dataset. Finally, the results are purely based on one property. Given the above, it is important to highlight that caution should be taken with extrapolating the findings in this report to new properties or scenarios.

The impact of the limited data used is more significant for hot water usage at the larger time horizons as for hot water usage cumulative consumption is used, whilst occupancy is not cumulative rather at a point in time. As such, hot water usage is very auto-correlated and there are very limited changes of class through a 7 day period with a 72 hour time horizon for the predictive model. Table 4 displays how the limited data affects both the performance of the predictive model and the decrease in performance when moving from the training data to the testing data. It is clear that the shorter time horizons perform far better and the longer time horizons perform worse and are more volatile. As such, future iterations of this work should explore increasing the data size to have a better longer term predictive model.

Time horizon	Predictive performance [AUC]	Performance on test data vs training data
10 mins	92%	2%
1 hr	70%	-5%
4 hrs	62%	-3%
24 hrs	61%	-10%
72 hrs	46%	-53%

Table 4: Hot Water Usage predictive model performance & volatility

3.3 Predictive model

3.3.1 Choice of machine learning model

The predictive models used for hot water usage and occupancy predictions required a few key characteristics:

- classification algorithm which outputs class probabilities;
- able to effectively use high dimensional input data;
- able to learn non-linear relationships leading to improved performance;
- provide feedback on the predictive power of different data features;
- have high predictive performance.

A Random Forest Classifier was chosen, given that the predictive problem was framed as a classification problem (low, medium and high values for hot water usage; occupied, not occupied states for occupancy) and that Random Forest Classifiers are known to have all of the above properties. The specific implementation that was used is the one provided with the Scikit-Learn library, which not only provides an expected class but also the probability of that class. Alternative algorithms could be considered in future research, potentially leading to improved performance.

It is important to note that the algorithm is intended to help predict the class that will appear as well as a rough probability of that class, but does not go as far as predicting the optimal action a HEMS system should take based on this information. To decide on an optimal action a fair bit of additional work is required, including calibrating the model probabilities, understanding energy prices, searching for an optimal solution, amongst others.

3.3.2 Formulating the right predictive problem

The overall exam question aimed at predicting residents' energy needs and three key areas were identified here, these are, hot water usage, occupancy and heating. The work focused on the first two. The reason for not building a heating needs predictive model is that heating needs are not directly observable from the data, instead it is only possible to observe the heating actions taken by the current control system. Heating control system signals are quite volatile making it quite hard to decouple the control systems behaviours from the human behaviours. For this reason, it was felt that the occupancy is a better proxy of heating needs, than the heating system's actions.

The predictive problem could be framed in many ways, each with advantages and disadvantages. For both predictive models, it was decided that models would be built for 5 different time horizons (10 mins, 1 hour, 4 hours, 24 hours, and 72 hours). The time frames were chosen using an understanding of the mechanics of heating a property and water, as well as an understanding of human behaviour, with the aim of picking a broad range of time frames in which any insights are actionable through the HEMS. In the case of occupancy, the predictive model was formulated in a way that it estimates the probability of occupancy (a binary value) at the end of that given time horizon. For instance, the 1 hour time horizon estimates the probability someone is home in exactly 1 hour, and is not concerned with predicting the time of occupancy between now and that hour. In the case of hot water usage, the model aims to predict the level of hot water usage over the given time horizon. For example, in the 1 hour model, the model would predict the likelihood that total hot water usage is high over that 1 hour time frame.

Alternative formulations of the predictive model were considered, which include:

- **Hot water usage quantity**: for hot water usage a classification model was created, in which the algorithm had to predict which type of usage would occur (low, medium or high). An alternative, would be to predict the amount of usage over the time period. The discretised version was chosen as the results are easier to interpret and the states could be broadly associated to different boiler setting options. That said, both are very reasonable formulations.
- **Time to next event**: an alternative formulation is to predict the time till a given event as opposed to the quantity at or over a given time. For instance, in how long will the property be occupied, as opposed to how likely is it to be occupied in 1 hour. The relative performance of the two formulations will be largely driven by the shapes of the distributions of hot water usage or occupancy, whilst the usefulness of the two approaches will depends on the specific of the HEMS system. It is worth exploring the alternative formulation here.

For the hot water usage model, the quantity that is ideally predicted is the total volume of hot water used at a given temperature. That said, this is not directly inferable from the dataset, and as such was approximated as the total number of seconds for which the hot water was on in that period. This does not explicitly differentiate between a trickling tap and full power water usage, but was felt to be a good proxy. The total time of hot water usage was then discretised into 3 categories for each time horizon. The thresholds were determined using human judgment that combined a physical understanding of the different time periods with data exploration to understand the distribution of usage over those time periods. The table below displays the different thresholds chosen as well as their approximate frequency, based on a small sample.

Time Horizon	Time range of medium hot water usage [secs]	Frequency of low hot water usage	Frequency of medium hot water usage	Frequency of high hot water usage
10 mins	**30-180**	**96%**	**3%**	**1%**
1 hour	**30-180**	**87%**	**8%**	**5%**
4 hours	**30-180**	**63%**	**17%**	**20%**
24 hours	**120-600**	**7%**	**25%**	**68%**
72 hours	**1500-4000**	**3%**	**81%**	**16%**

Table 5: Hot water usage thresholds and frequency of occurrence

3.3.3 Feature selection and calibration of model

The approach followed to determine the ideal combination of features and calibration of the model hyper parameters was slightly different for the hot water usage and occupancy predictive model, as such they will be explained separately. It should be noted that the number of combinations of features and hyper parameters are too great to make a full search realistic, and as such a local maxima rather than a global maxima is obtained.

For the hot water usage model, the features were selected by working in two directions. In one case, a large list of features was included and the least predictive were removed, reaching a more powerful combination of features. In the second case, the most predictive feature was included first and features were added to increase the predictive performance. Similar models are obtained in both cases and compared, to obtain a best model. Whilst the main driver for adding or removing a feature was the feature importance score obtained, the AUC was monitored in order to ensure that the model was improving.

For the occupancy model, the features were selected by incrementally adding groups of features, and after adding each group testing whether the AUC performance increased for the cross-validation set. The groups of features included all contributed a material difference to performance so all features were included.

Random Forests have various hyper-parameters that can be calibrated to avoid overfitting and increase predictive performance. Parameters considered were max depth, minimum sample split, number of trees and minimum sample leafs. The latter parameter was searched for an optimal solution as its predictive power is less volatile making it easier to find the optimal parameterisation and this also ensures performance is more stable. For both hot water usage and occupancy the min sample leaf was optimised for by running a range of values and observing where the AUC peaks. The number of trees was optimised more loosely, by increasing the number of trees to a level that was computationally tractable, resulting in 500 trees for all occupancy models and 500-1000 tress for all hot water usage models.

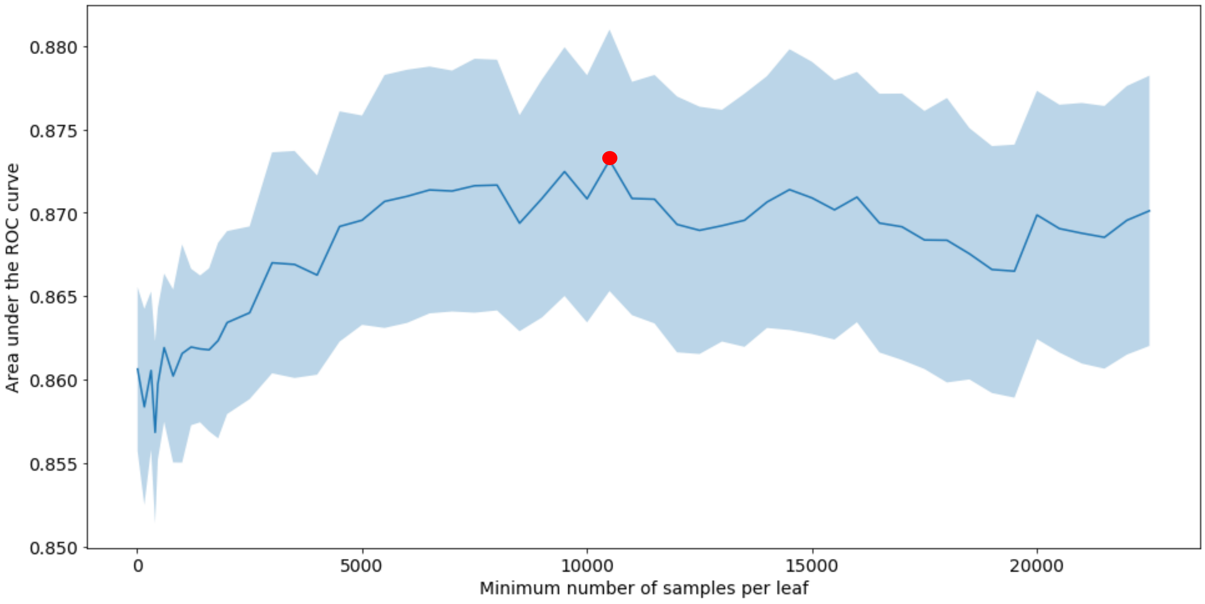

Figure 2: Parameter search for optimal minimum number of samples per leaf using occupancy at a 1-hour time horizon with 500 trees. A value of 10,500 was found, which could be further optimised by searching more granularly, and was applied to all 5 occupancy models.

4 Feature engineering

Data feature engineering is a key area of focus when developing a machine learning solution as it allows for the domain expertise to be captured mathematically, often leading to significant increases in predictive performance. This problem required feature engineering for two main reasons. The first is that the problem captures 615k data points a second leading to a multiple TB data size challenge. In order to make the problem computationally tractable it is necessary to apply data compression in a way that minimises information loss and reduces any impact to predictive performance. Additionally, this problem requires feature engineering to enhance predictive performance, as do most machine learning problems. The idea here is that humans have a good understanding of human behaviours and these should be explicitly captured in the data. For instance, we know that water usage is often cyclical at the 24 hours level and as such it should be captured through a data feature using 24 hours of memory.

4.1 Data compression

Data compression was only required for the electricity data as this dataset was at 205kHz, whilst the other datasets were all at less than 1Hz. In order to determine a suitable approach to data compression, some data exploration was carried out and domain knowledge and previous research were brought to bear. The overall approach chosen was one in which the data was first reduced by a factor of 75 through a technique we refer to as peak finding, followed by a further data size reduction of 160 through Principal Component Analysis (PCA).

4.1.1 Data exploration

In order to inform our data compression strategy, data exploration was conducted on the electricity data. Electricity was recorded at 205kHz as two current signals and one voltage signal. Given that the signal recorded was at 50Hz there were around 12,000 readings for any one sinusoidal wave of electricity. It was felt that the key information could be represented with far less data points, but it was not evident what these data points may be. For instance, it may be that apparent power and phase angle hold all the key information, or that changes in power are key, or that the first 10 harmonics are key. Whilst it would be ideal to test all the various theories by measuring the predictive performance of the algorithm, this was not realistic in the given time frames, and as such it was necessary to make such decisions purely through data exploration and human judgement.

A key observation made through the data exploration conducted is that there is high intensity of frequencies at multiples of 50Hz, as per Figure 3, which led to a harmonics driven down sampling approach. It can also be observed, as per Figure 4, that odd harmonics appear to contain more power than even harmonics and to have more consistent phase angles. An alternative approach to down sampling that was explored is a low pass filter up to 16kHz or similar. Figure 5 suggests that on the sub-sample of data analysed a good reconstruction is achieved using this approach and Figure 6 suggests that there are very few peaks above 5kHz. The approach was steered away from as whilst it would reconstruct this signal well, the impact on the predictive performance could not directly be measured and it was felt important to use the full frequency spectrum.

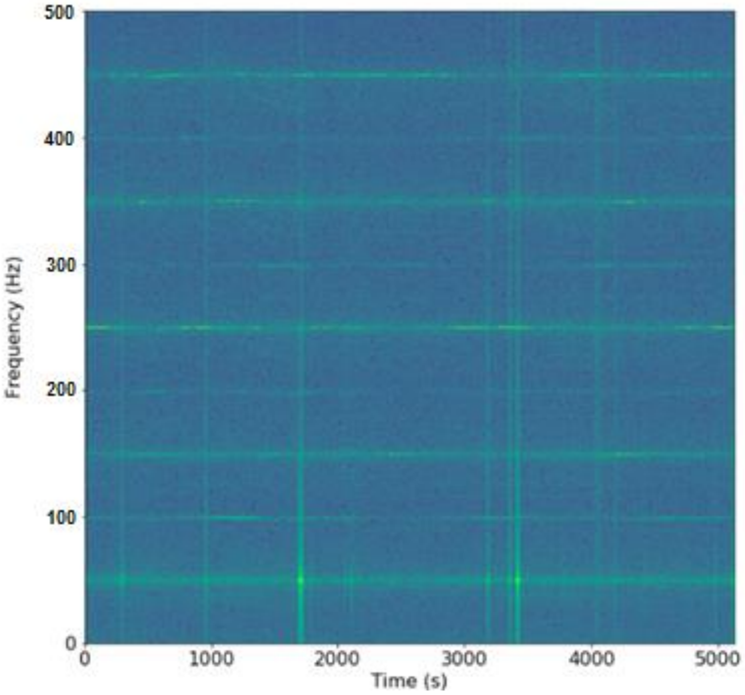

Figure 3: Spectrogram of the intensity of different frequencies present when applying approximately 5000 Fourier Transforms, one per second of data. Figure demonstrates that there is high intensity at the multiples of 50Hz, informing our approach to down-sampling, which is harmonics driven.

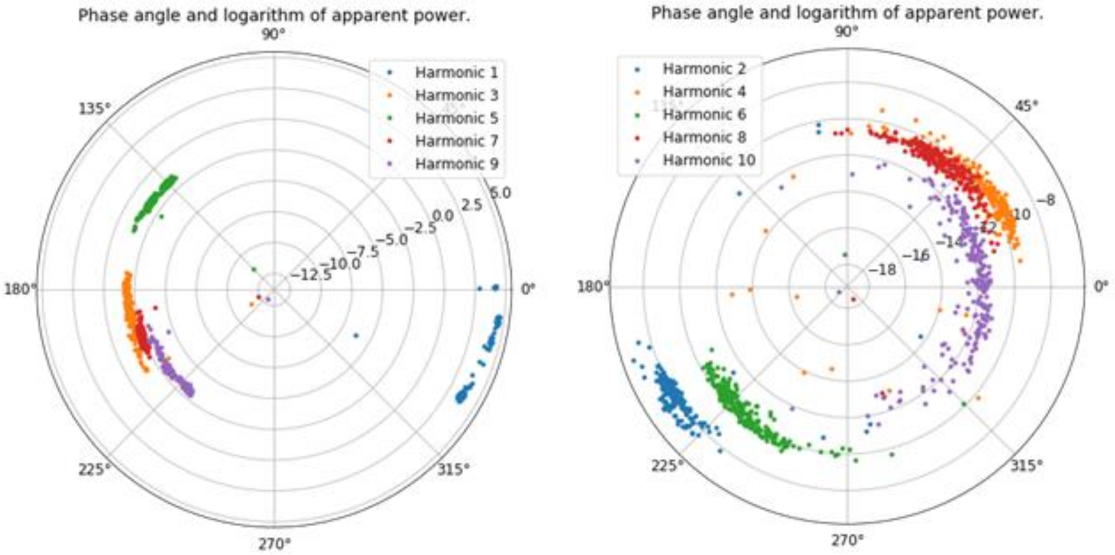

Figure 4: Each point represent the phase angle and the logarithm of apparent power for a given harmonic at a specific point in time. Odd harmonics contain more power than even harmonics, and have more consistent phase angles

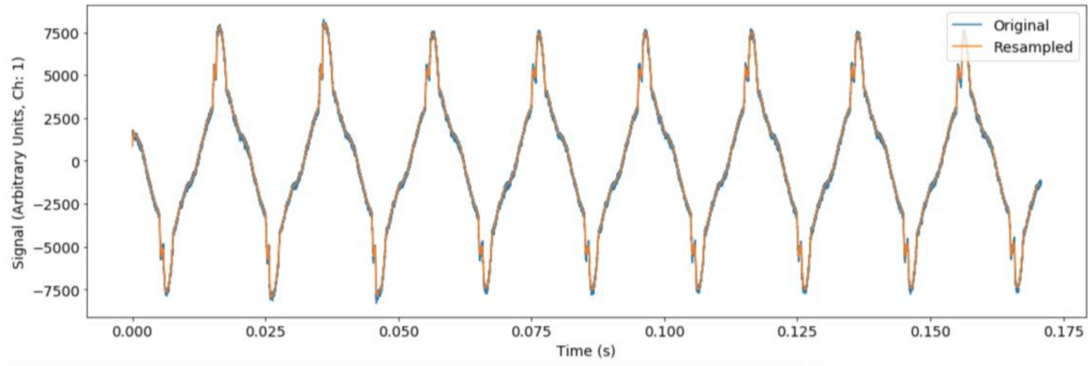

Figure 5: Illustration of the reconstruction of a sample of current signal using the Fourier Transform up to 16kHz and removing any data above that frequency, demonstrating that a good reconstruction is achieved. That said, the predictive value of the missing data is unknown. It was therefore decided to keep the full 205kHz spectrum, but to compress through lower resolution

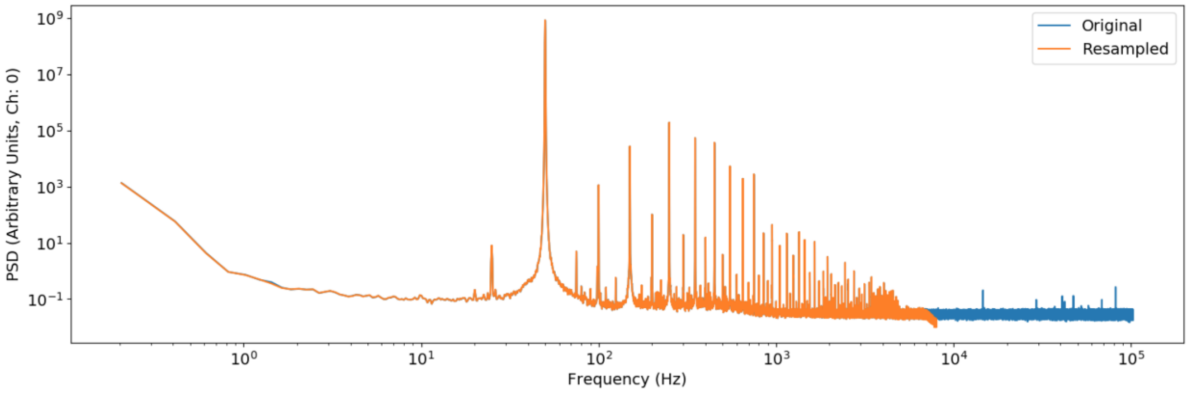

Figure 6: Spectrogram of the frequencies present in a sample of voltage data. Evident that intensities peak at multiples of the fundamental frequency (~50Hz) and that above 5000Hz there are very few peaks.

4.1.2 Fourier Transforms

Leveraging the data exploration in section 4.1.1, it was decided that the full range of frequencies would be kept for analysis, and reduced in data size by only storing the peak intensities of active and reactive power at the harmonics. This method allows for a reduction in data size of 75x as only 8000 data points are stored for every second of data, whilst still capturing the full breadth of frequencies. Below we share the specifics of the process used:

- Use the current signal, not power, as this significantly reduces the computational power required and the voltage signal is thought to have significantly less information than the current signal.
- Two current signals are stored in the raw data with varying resolutions. For any given second, if the min/max of the more precise current signal, which has a smaller value range, is outside a given range then the coarser signal is used, if not the finer signal is chosen.

- Calculate the Short-Time Fourier Transform (STFT) on the signal over a rolling time window.
- From the spectrogram there were two options on how we store the data at the various harmonics, peak binning (integrating over the range of c. 50Hz) or peak finding (storing the peak value). Both methods were felt comparable, but due to computational time being c. 3x quicker for peaking finding, peak finding was chosen. The approximate cost to process 1 month of electricity data for 1 property was £30.
- Finally, we evaluate different metrics at the harmonic peaks (active power, reactive power, phase angle, apparent power, complex amplitude of voltage, etc.), two of which are key in the analysis: active power and reactive power.
- A key design consideration is the size of the stride and the time window applied. Smaller strides provide more detail on human activity, at the expense of a smaller level of data compression and worse resolution in frequency. Option 1 in the table below was chosen, which has a frequency resolutions of ±0.16Hz (Δf = 1/(2πΔt)).

Option	Stride size	Window size	Frequency resolution [Hz]	% of server RAM required for processing (32GB)
1	1 second	2 seconds	0.16	90%
2	2 seconds	4 seconds	0.32	90%
3	5 seconds	7 seconds	0.80	61%

Figure 7: Potential stride and window sizes for the STFTs

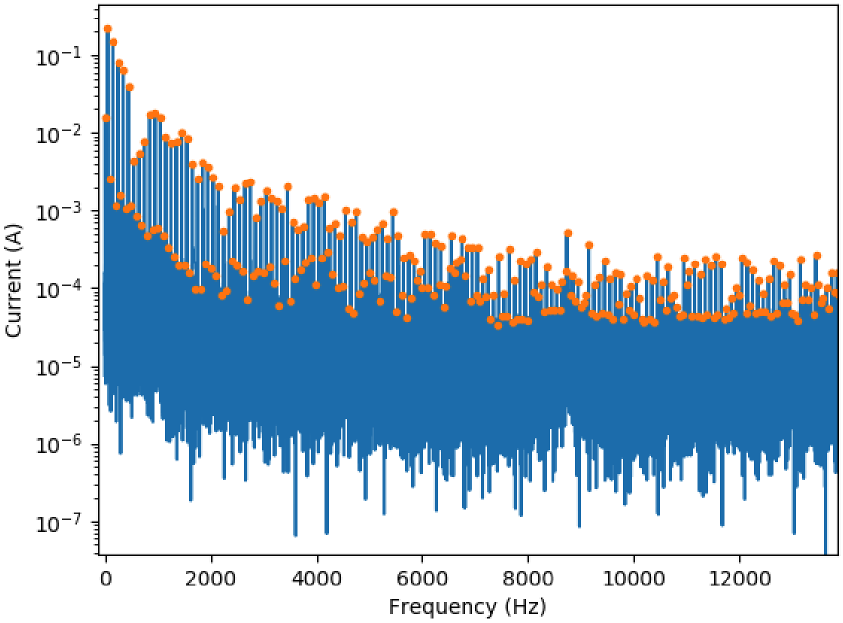

Figure 8: Peak finding applied on spectrogram and peaks visualised through orange circles

4.1.3 Principal components

Following the peak finding exercise, which reduced the dimensionality by 75x, there were still 8000 data points in a given second. In order to reduce this further, Principal Component Analysis was applied to the peak finding data, and the 50 main principal components stored, reducing the data by a further 160x. The reconstructive power of the first two principal components is significant, see Figure 9, making an argument for only keeping those two if the objective was to reconstruct the electricity signal whilst maintaining high levels of data compression. However, the end goal is to predict occupancy and hot water usage, and it was therefore decided to keep more Principal Components to test if they had strong predictive power. The results of the predictive models suggest that the first three principal components all have significant and comparable predictive power, with some other principal components faring quite well on occasion. Figure 10 visualises the first 5 Principal Components, making their physical meaning more apparent.

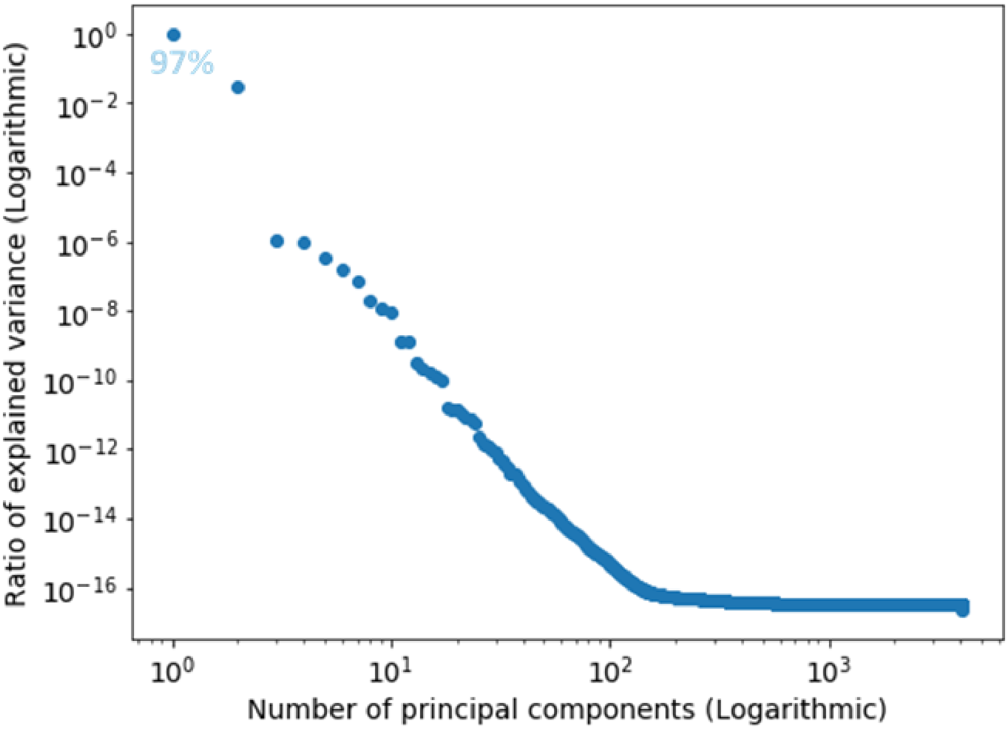

Figure 9: Reconstructive power of first 50 Principal Components

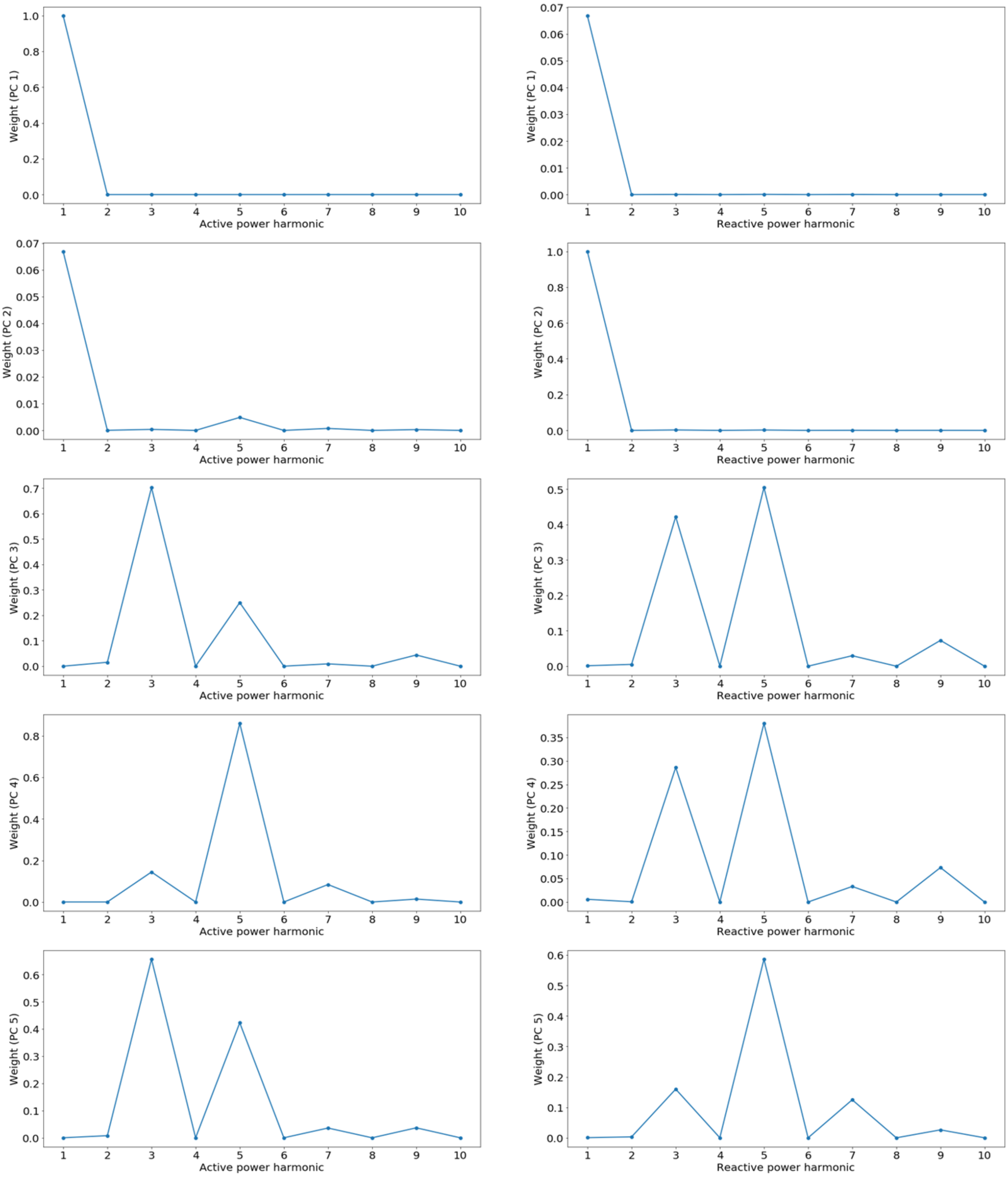

Figure 10: Visualisation of 5 main principal components by the energy present in the first 10 harmonics for both active and reactive power. PC1 appears similar to active power, PC2 appears similar to reactive power and PC3 is heavily focused on the third and fifth harmonics.

Non-linear techniques like UMap were considered and implemented, but due to the data size, it was preferable to use a dimensionality reduction technique that is incremental and parallelisable. IPCA proved to be one of the few options available and as such, the project steered away from more mathematically advanced techniques such as UMap.

4.2 Data feature engineering

Dozens of data features were engineered and tested to increase predictive performance. Here we break these down into a few categories: electricity clusters, memory, and exogenous factors auto-regressive variable.

4.2.1 Electricity clustering

As outlined in section 4.1, it was necessary to downsize the electricity data through peak finding and principal component analysis. This provides a vector representation of every second of electricity data, but it does not exploit any form of clustering of the electricity patterns. With the purpose of understanding electricity behaviour, labelling occupancy and improving predictive power a clustering of the compressed electricity data was run. This led to the grouping of electricity behaviour into 25 categories that represented 89% of the data. These clusters were then used in the predictive algorithm as 25 binary features. Although the clusters had predictive power, it appears that the algorithm is able to obtain most of that performance by directly mapping the electricity principal components to the output, suggesting that the PCs should not be removed in place of the clusters. However, future work on the clusters ordering through techniques such as correlation analysis or natural language processing, is likely to yield a better understanding of human workflow and introduce a new type of memory data feature that should yield performance increases. It is also important to note that the clusters were key as an intermediary step to label the property's occupancy state.

For the clustering algorithm, HDBSCAN was chosen over techniques such as k-means as a visual inspection of the data suggested the clusters are not sphere shaped. Additionally, this specific algorithm is readily available in GitHub in a version that is parallelised. A specific challenge with clustering is the parameterisation of the function, which has no ground truth to evaluate performance against. In our case, we varied two parameters (min cluster size & min samples) in order to determine a parameterisation that gave a visually sensible result. The results chosen based on running the analysis on property H45 is shown in Figure 11.

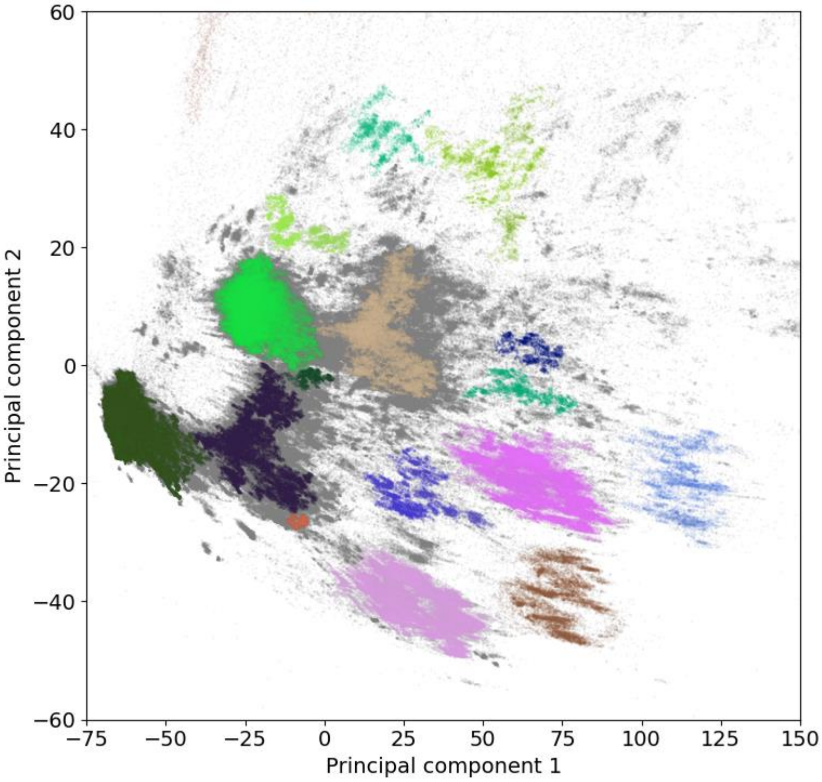

Figure 11: Results of running final parameterisation of HDBSCAN on 75 days of H45 data. Analysis run on 50 principal components and visualised for top 2.

4.2.2 Memory

Memory was introduced in many forms and it appeared to be critical to achieving high predictive power. The specifics of the way it was included depended on the variable it was being applied to and the timeframe over which the predictions were being carried out. Key representations of variables include:

- The consecutive usage, time of usage or absence of a variable i.e. the consecutive absence of hot water usage;
- Changes to the state over a period of time i.e. increment in humidity;
- Consumption over a period i.e. hot water usage over a period of 24 hours;
- Exponentially weight moving average (EWMA) to include water consumption or electricity usage over a time range with a stronger focus on the more recent consumption. It is often beneficial to include multiple time frames for a given model, and it appears that longer time frames work better for predictions over longer time periods.

Further work could have been done in creating specific memory features and linking these to known human activities. For instance, the toilet flush seemed to consume 5.5Liters and refill at 110ml/sec. This could be captured in the data features by going beyond features capturing cumulative consumption over a time frame and including features such as total number of times consumption appeared, the lengths of those consumptions, amongst others.

4.2.3 Exogenous factors

A few exogenous factors were introduced to the master dataset to improve predictive performance and act as a proxy of priors. These included:

- Time of day;
- Weekday;
- Non-working day;
- Light/dark based on sunrise and sunset;
- Mealtime.

The most predictive prior was generally time of day, which is not surprising. In isolation it holds relatively strong predictive power, which means that if the system were to have data collection issues for a period of time, the system could revert to predictions using the prior proxies. There is also the potential to carry out further work on priors, for instance week days could be grouped differently, or using month of year once more data has been prepared.

4.2.4 Auto-regressive terms

Using the current state and historical states of the target variables proved to have strong predictive power i.e. current hot water usage is a good indicator of future hot water usage. As such, future solutions should consider ways to introduce these variables into the predictive model. In the case of hot water usage, this should not present significant complication, as it is a directly observable variable. In the case of occupancy, the results assume that it is possible to observe occupancy and that this is stored. For this specific setup that is not the case, but the objective here was to test the art of the possible. Future systems will need to consider whether they somehow observe occupancy, producing comparable results to the results reporting here, or whether the predictive model uses a proxy such as predicted historical occupancy.

4.2.5 Performance of various feature categories

It is important to understand the predictive value of different features and categories of features when considering a future system of this sort. This would allow for an informed decision on the equipment needed so as to maximise performance and keep cost down. Whilst understanding what the ideal production ready solution would look like is out of scope of this analysis, in developing the algorithms a broad understanding of the predictive power of different feature types was obtained. Taking the example of the 1 hour occupancy predictive model, we observed the following:

- First generation model:
    - 69% Area Under the Curve (AUC).
    - 4 of the top 6 features were exogenous features i.e. hour, day of week, mealtimes and bedtime.
    - 2 of the other top 6 features are bathroom humidity and pipe hot water temperature.
    - The clusters provided low additional value and the top 3 were those that were labelled as autonomous.
- Second generation model:
    - Model now includes first three electricity principal components leading to an increase in performance, now obtaining 75% AUC.
    - Three principal components have comparable predictive performance, with the importance being the inverse of what may be expected i.e. PC3 has the strongest performance and PC1 has the lowest.

  - Around half of the increment in performance is due to the addition of the principal components, the other half due to memory in those principal components. Short-term memory is introduced through exponential weighted moving averages.
- Third generation model:
  - Auto-regressive terms are included, historical and current occupancy, leading to a large jump in performance to 88% AUC.

5 Hot water usage modelling results

Five predictive models were built for hot water usage, one for each of the agreed time horizons. This section is broken down into 5 sub-sections, one for each time horizon, and in each sub-section the key predictive features as well as the predictive performance are shared.

As may be expected, the performance drops as the time horizon increases, as per Table 6. Additionally, it is evident that the stability of the solution deteriorates with increasing time horizons, as measured by "performance on test data vs training data" in Table 6. The extent of the deterioration at 24 and 72 hours, makes a strong case for the need to extend this analysis to a larger dataset and points out the need to be careful in generalising too far from the small dataset that has been used.

In terms of feature importance, the key features are consistently hot water usage (auto-regressive feature), water usage and bathroom humidity, with electricity principal components having some influence at different time horizons. The memory of the data features was also a key input with memory of comparable life to the predictive time horizon and 24 hour memory generally performing well. Interestingly, electricity clusters did not feature well and exogenous factors had a minimal influence.

Time horizon	Predictive performance [AUC]	Performance on test data vs training data
10 mins	92%	2%
1 hr	70%	-5%
4 hrs	62%	-3%
24 hrs	61%	-10%
72 hrs	46%	-53%

Table 6: Hot Water Usage predictive model performance & volatility

5.1 10-minute time horizon

The hot water usage model performs significantly better on the 10-minute time horizon as compared to any other time horizon. As one may expect, the key predictive features relate to water usage, the temperature of the pipe used for hot water and bathroom humidity. The overall performance is 92% AUC on the test set, with performance being strongest for the high hot water usage classifier. **Note**: the random forest classifier predicts a probability for each of the three classes, but the AUC plots have been generated by treating each class individually as a binary classifier. For instance, in the case of class 2, high hot water usage, the performance of the classifier is measured for predicting high hot water usage vs low & medium hot water usage grouped together.

	Predicted Low	Predicted Medium	Predicted High	Total
Actual Low	91.2%	5.2%	0.2%	96.6%
Actual Medium	0.8%	1.3%	0.3%	2.4%
Actual High	0.1%	0.4%	0.5%	1.0%
Total	92.1%	6.9%	1.0%	100%

Table 7: Confusion matrix for hot water usage prediction at a 10-minute time horizon. In this example, the model has been calibrated to over-predict medium cases, as this is hard to identify and it manages to capture over half of the actual cases despite medium usage only occurring 2.4% of the time.

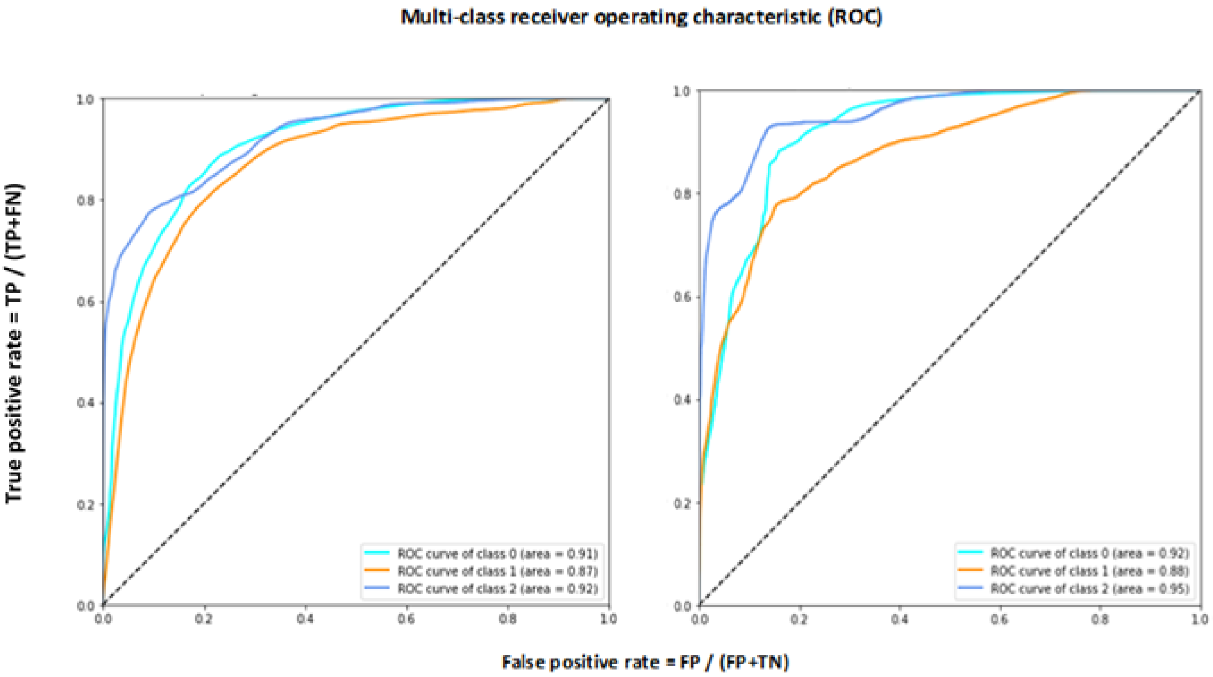

Figure 12: Hot Water Usage predictive performance for 10-minute time horizon. **LHS**: cross-validation results. **RHS**: testing results. Class 0 is low hot water usage, class 1 is medium hot water usage and class 2 is high hot water usage.

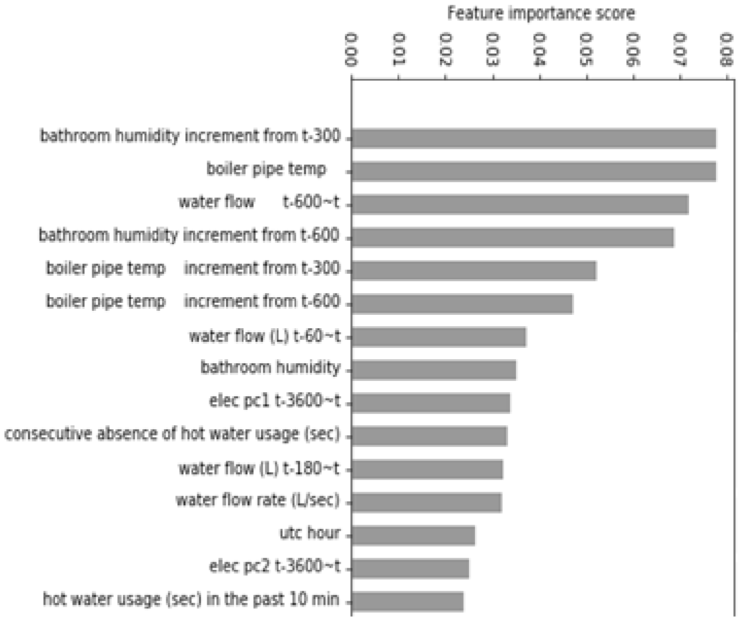

Figure 13: Random Forest ranked feature importance for 15 most powerful features in 10 minute hot water usage predictive model

5.2 1-hour time horizon

The hot water usage has a 75% AUC on the training set and a 70% AUC on the test set. The significant drop in performance from the 10-minute time horizon is largely down to it being more difficult to predict for a longer time-frame, but is also likely due to the effective training data being 6x smaller. The lack of independent 1 hour time periods to train the data on, make the problem more prone to overfitting and reduces performance, as observed with a larger drop in performance between the training and testing sets. In terms of the features that perform well in this time horizon, these are the first 10 electricity principal components using memory with a 2 hour half-life (comparable magnitude to the 1-hour prediction time horizon) as well as water usage in the last 21, 23 and 25 hours. In contrast, electrical principal components beyond the first 10 and the electricity clusters did not fare so well.

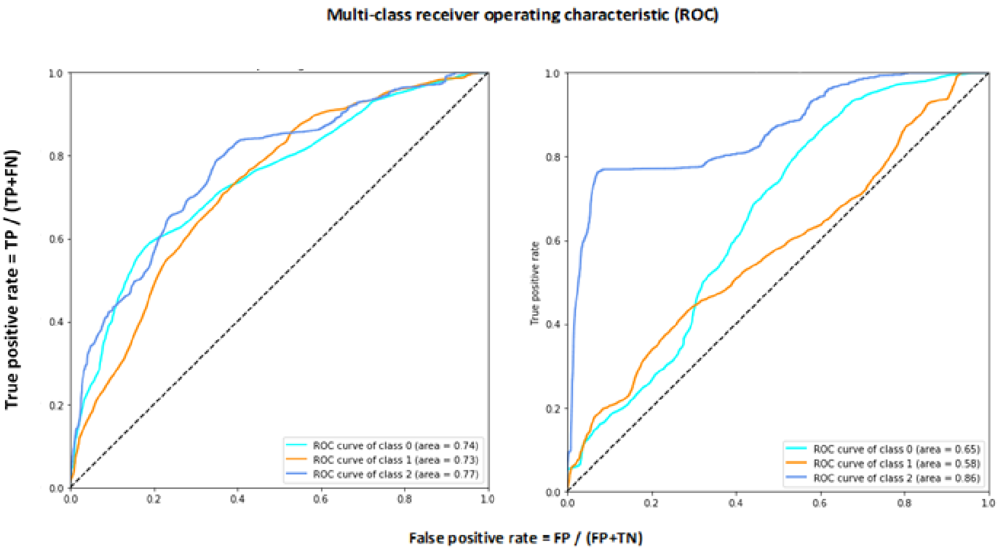

Figure 14: Hot Water Usage predictive performance for a 1-hour time horizon. **LHS**: cross-validation results. **RHS**: testing results. Class 0 is low hot water usage, class 1 is medium hot water usage and class 2 is high hot water usage.

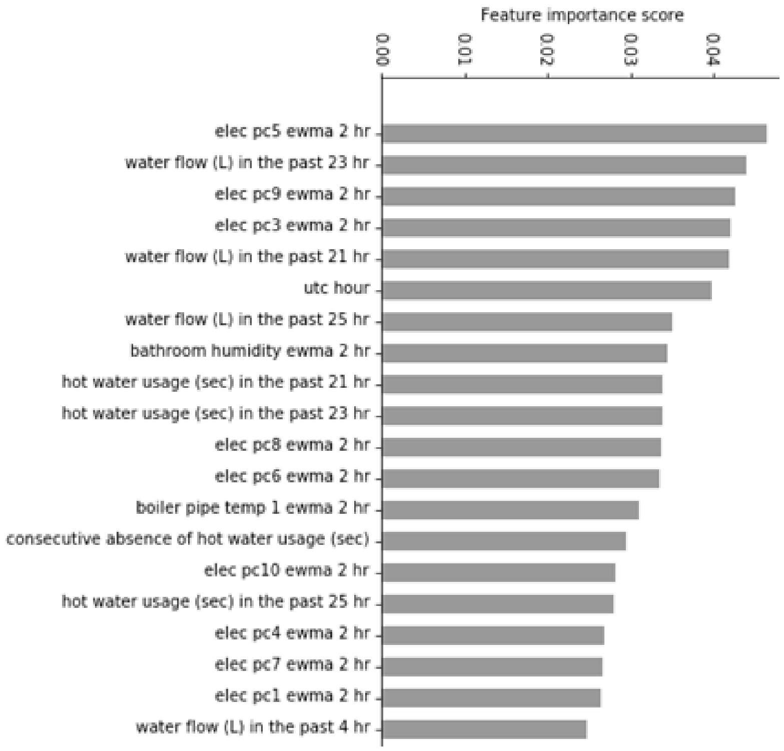

Figure 15: Random Forest ranked feature importance for 20 most powerful features in 1-hour hot water usage predictive model. "EWMA x hr" stands for exponentially weighted moving average with a half-life of x hours.

5.3 4-hour time horizon

The hot water usage has an average of 65% AUC on the training set and 62% AUC on the test set. The volatility in performance between the training and testing sets is disguised by taking the average and is significant as can be seen in Figure 16, suggesting that there is a key behavioural change or a shortage of data for the current models. In terms of the predictive variables, water usage and hot water usage carry significant weight as per Figure 17.

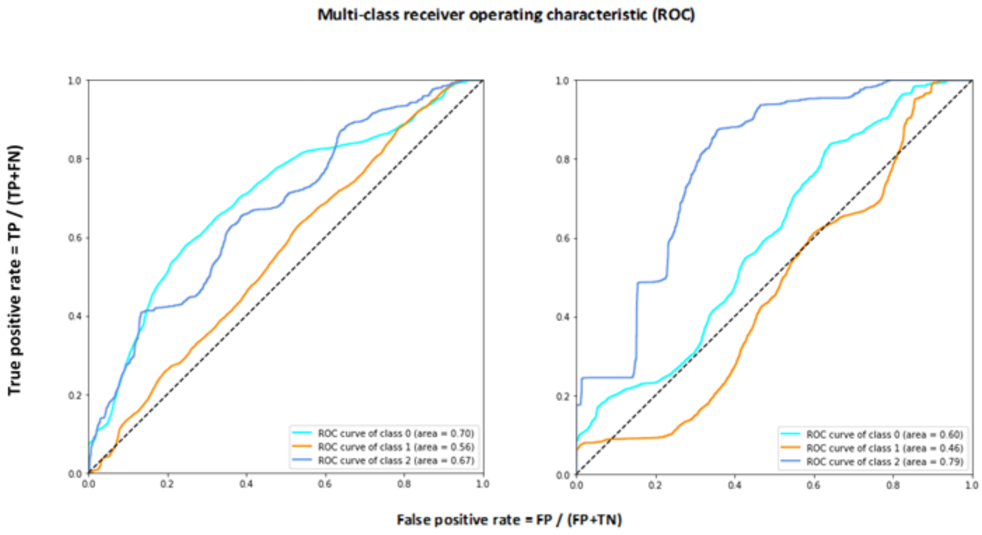

Figure 16: Hot Water Usage predictive performance for a 4-hour time horizon. **LHS**: cross-validation results. **RHS**: testing results. Class 0 is low hot water usage, class 1 is medium hot water usage and class 2 is high hot water usage.

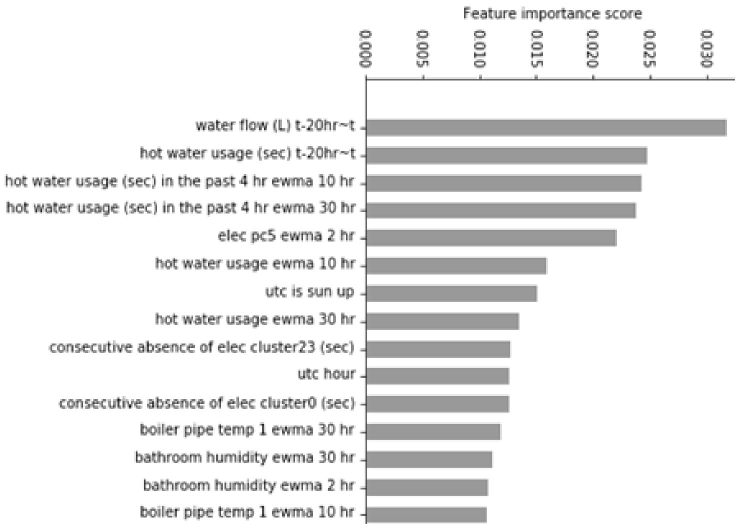

Figure 17: Random Forest ranked feature importance for 15 most powerful features in 4-hour hot water usage predictive model. "EWMA x hr" stands for exponentially weighted moving average with a half-life of x hours.

5.4 24-hour time horizon

The hot water usage has an average of 71% AUC on the training set and a 61% AUC on the test set. The volatility in performance between the training and testing sets is becoming significant, suggesting that there is a key behavioural change or a shortage of data for the current models. In terms of the predictive variables, water usage and hot water usage carry significant weight especially when over a time horizon with a magnitude close to 24 hours, as per Figure 18.

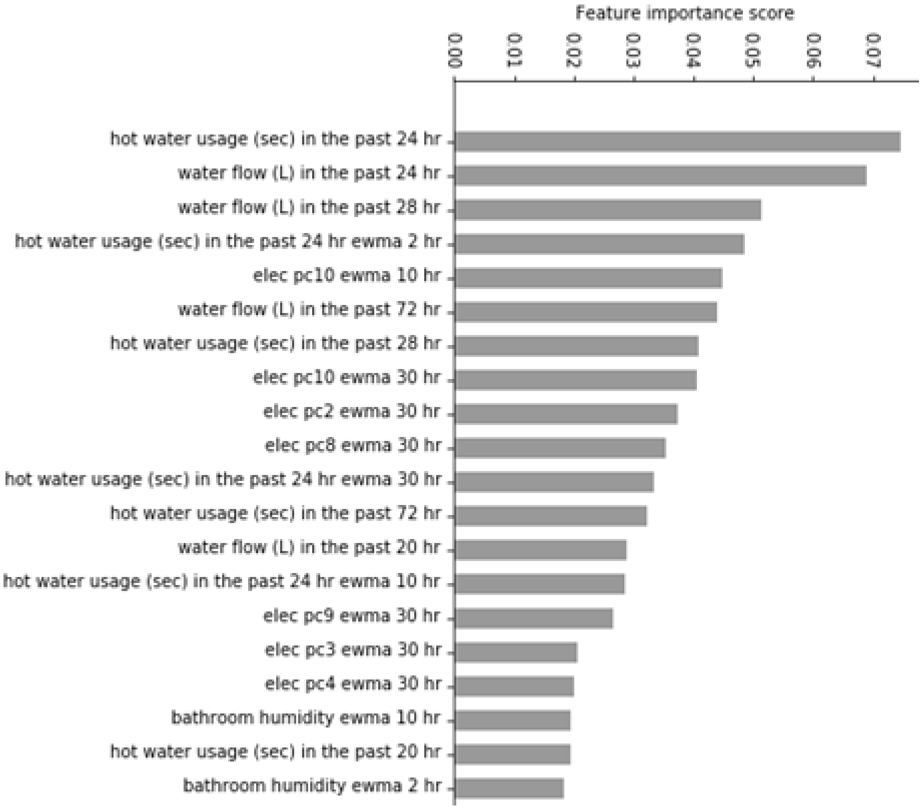

Figure 18: Random Forest ranked feature importance for 20 most powerful features in 24-hour hot water usage predictive model. “EWMA x hr” stands for exponentially weighted moving average with a half-life of x hours.

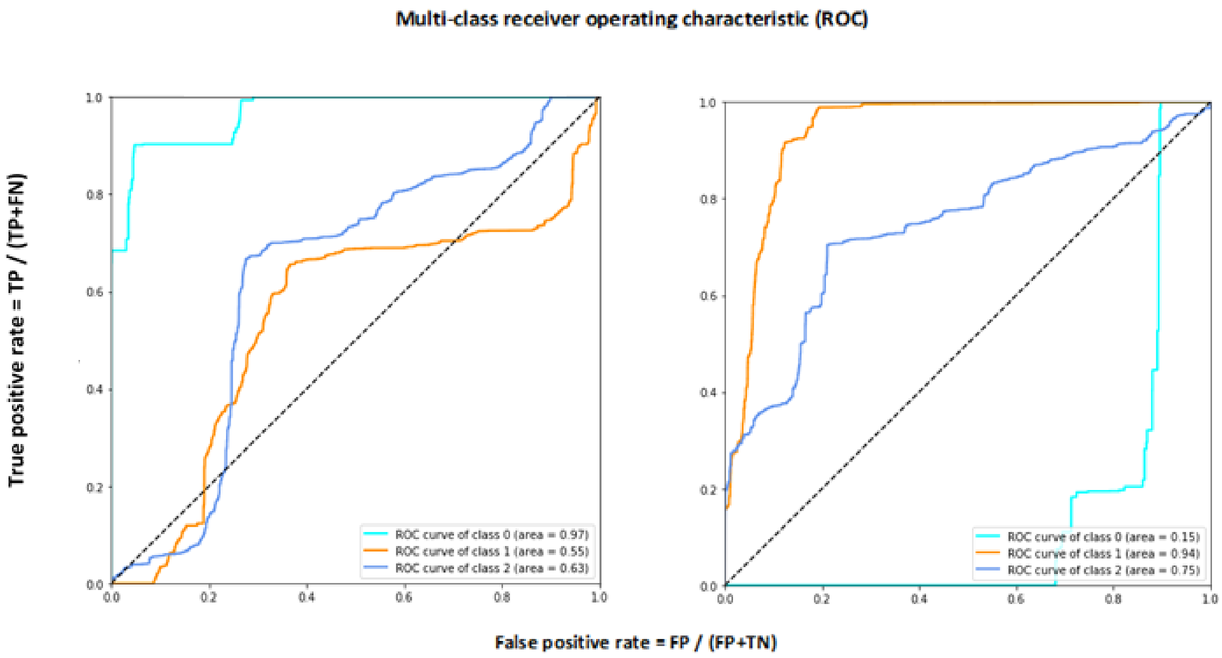

Figure 19: Hot Water Usage predictive performance for a 24-hour time horizon. LHS: cross-validation results. RHS: testing results. Class 0 is low hot water usage, class 1 is medium hot water usage and class 2 is high hot water usage.

5.5 72-hour time horizon

The hot water usage predictive model for a 72 hour time-horizon has limited value. Given that hot water usage is cumulative and the test set is 7 days, it could be argued that there are really only two independent data points to test on, as any other data point will be highly cross-correlated to the first two. That said, the results of training and testing are shared in Figure 20. In terms of the predictive variables, various data categories perform well including water usage, hot water usage, humidity and electricity. The key for high performance is memory or half-life of comparable magnitude to the time-horizon of 72 hours, as per Figure 21.

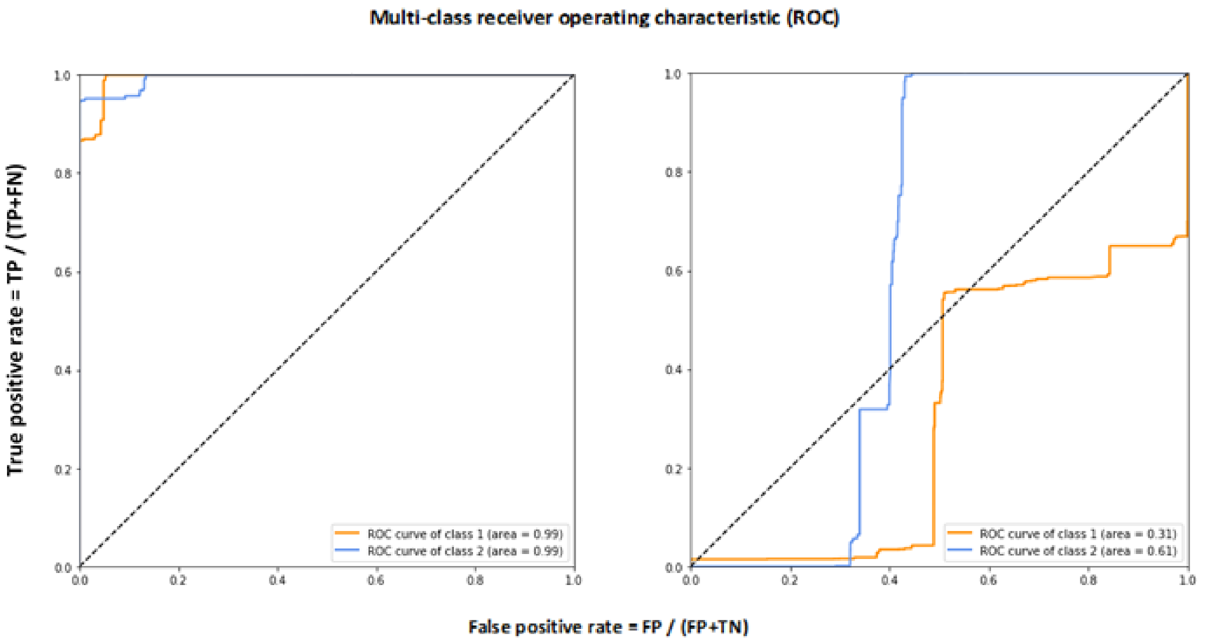

Figure 20: Hot Water Usage predictive performance for a 72-hour time horizon. **LHS**: cross-validation results. **RHS**: testing results. Class 0 is low hot water usage, class 1 is high hot water usage. Given that there are not many unique 72 hour groupings of the dataset, only two classes were used for the 72-hour time horizon model.

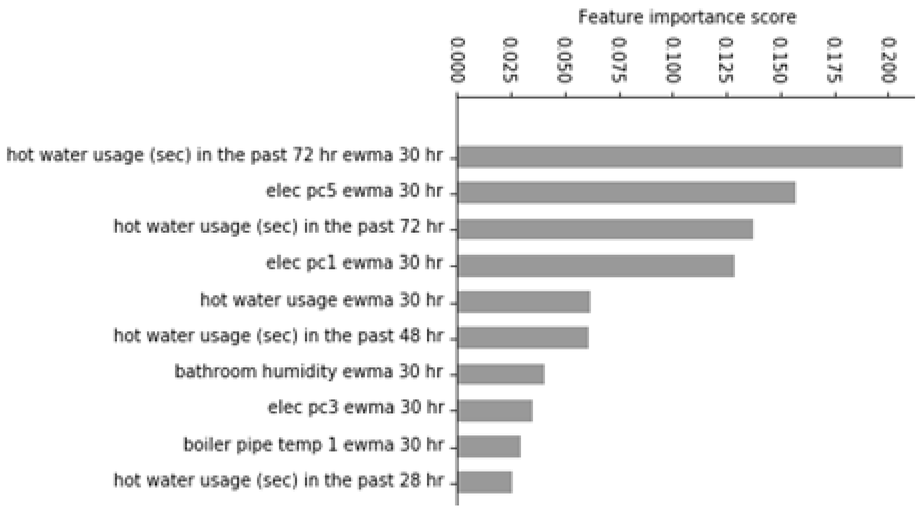

Figure 21: Random Forest ranked feature importance for 10 most powerful features in 24-hour hot water usage predictive model. “EWMA x hr” stands for exponentially weighted moving average with a half-life of x hours.

6 Occupancy modelling results

This chapter is broken down into two sections. The first section explains how the occupancy labels were generated, which is key in interpreting the significance of the results. The second section elaborates on the results obtained for the five different time horizons.

6.1 Labelling occupancy

In order to label occupancy, a two-stage process was followed starting from the electricity clusters detailed in Section 4.2.1. In the first stage, clusters were labelled as autonomous and manual, in the second stage the cluster labels and other data points were used to label occupancy.

6.1.1 Interpreting clusters

The first step in determining if someone was home was to profile the 25 electricity clusters, that represent 25 different property states, as autonomous or manual. In order to facilitate that labelling, a figure was produced for each cluster detailing the frequency of occurrence of that cluster for each hour in the 3 months of data used, as per Figure 22 and Figure 23. A few individuals looked over each of the clusters and labelled them accordingly.

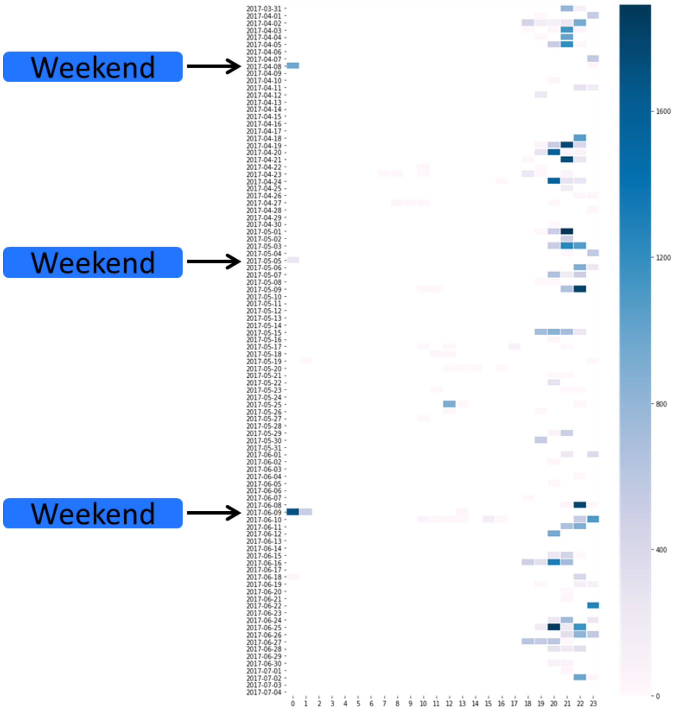

Figure 22: Frequency of a manual cluster throughout 24 hours in a day (x-axis) over a period of 3 months (y-axis). This cluster was labelled as manual as it only appears to come on in the evening and it is not extremely consistent i.e. reflecting sunset or a timer activated state**.**

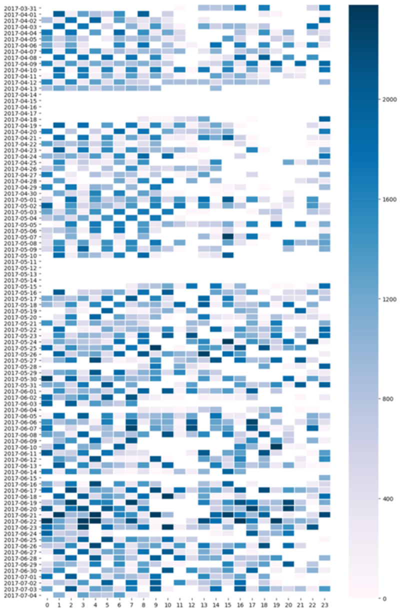

Figure 23: Frequency of an autonomous cluster throughout 24 hours in a day (x-axis) over a period of 3 months (y-axis). This cluster was labelled as autonomous as it appears through a broad range of time, suggesting it is not a human activated activity. The white patches are often due to data availability.

The end result of the labelling exercise is that 21 of the 25 clusters are labelled as manual and 4 as autonomous, with 11% of data points not getting a cluster label and being treated as between the two states by receiving a value of 0.5 in the mathematical representation (1 is autonomous, 0 is manual). As may be expected, the autonomous clusters occur very close to zero for the first two principal components, which are comparable to active and reactive power.

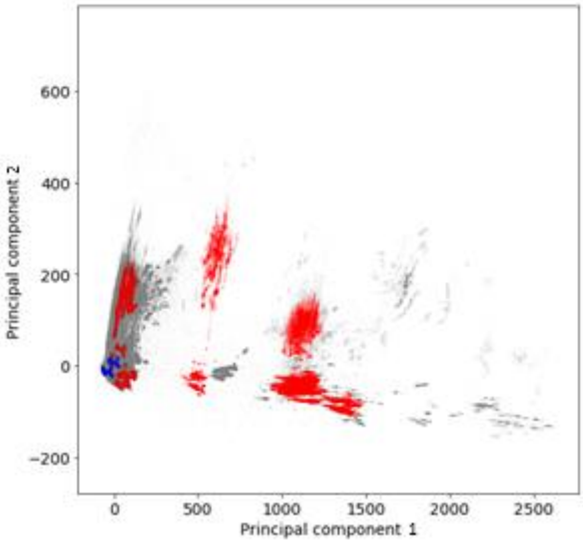

Figure 24: Autonomous clusters are blue, manual clusters are red, unlabelled data points are grey.

6.1.2 Occupancy label

In order to identify occupancy, a combination of data features are used including, a likelihood based on manual property states based on electricity clusters, historical water usage, temperature of hot water and time of day. Figure 25 provides an illustration of an application that was built to facilitate the occupancy labelling using a visualisation of the various data features. For every 10 minute period the user makes a judgment on whether they believe the property is occupied, and enters that into the app.

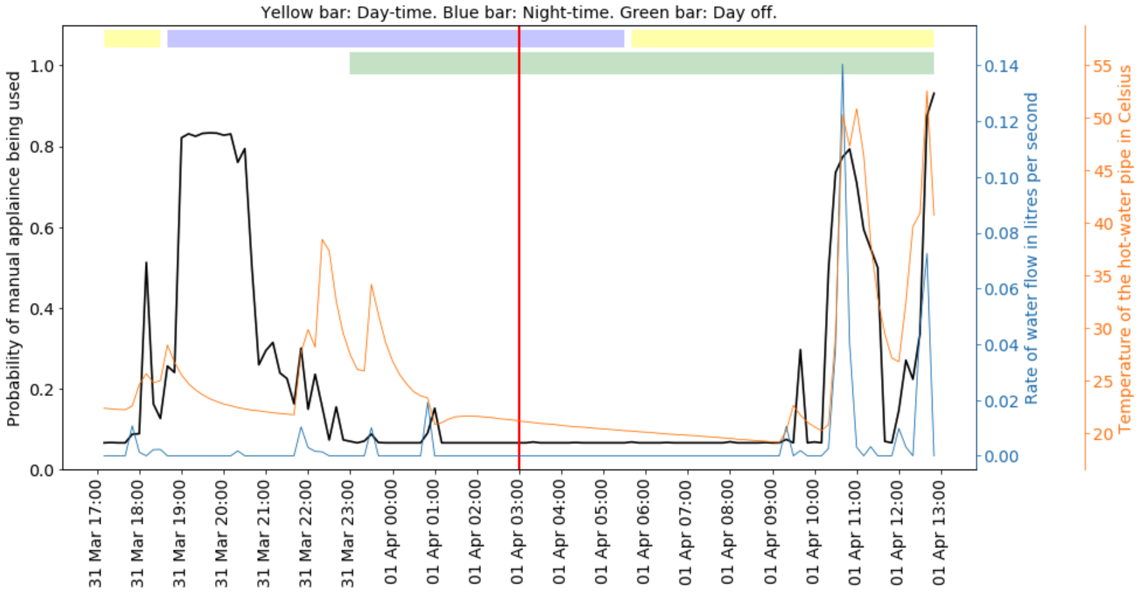

Figure 25: Application used to label property occupancy. The red line displays the current 10 minute interval we are evaluating for occupancy and the other data points are used to determine whether the property is thought to be occupied using visual inspection and heuristics

The list below ranks the factors that the data scientist examined when generating the label of occupancy from the most to the least important:

1. High probability of the HDBSCAN cluster being 'non-autonomous' was taken to signify that the house is occupied.
2. If activity was detected around bedtime and wake-up time of the next day, the hours in-between were labelled as 'occupied', regardless of the night appearing to be quiet. It was assumed that the occupants prepared for sleep in the evening, rested, and got up in the morning.
3. Mid-to-high consumption of hot water was deemed to imply that the house is occupied.
4. Breakfast, lunch and dinner were analysed with a mild expectation that occupants are at home. For example, the occurrence of manual clusters was interpreted as cooking. Lack of activity around mealtimes, on the other hand, was considered to indicate that the house is empty.
5. Periods of quiet that lasted more than 15 hours, especially before a weekend or a holiday, were attributed to the occupants being out of town.

The 39 days of H45 data were labelled in 10-minute increments as described. The result of this is a property that is thought to be occupied 74% of the time, with the property typically being vacated twice a day. The full view of the occupancy over the 39-day time window is displayed in Figure 26.

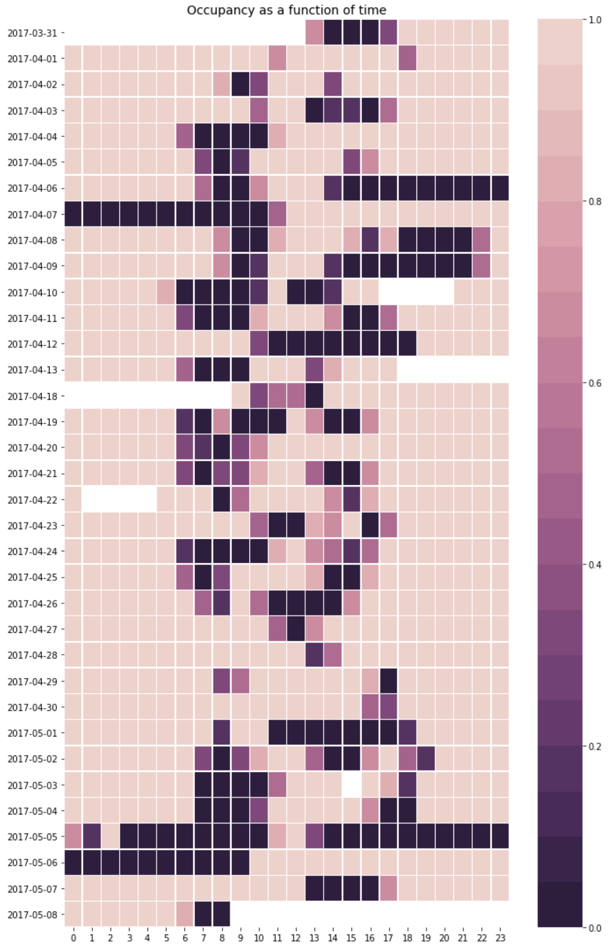

Figure 26: Heat map of occupancy of the property for a given hour. The x-axis displays the 24 hours in a day and the y-axis displays the 39 day period analysed.

6.2 Results

This section presents the results for the 10-min, 1-hour, 4-hour, 24-hour and 72-hour occupancy predictive models. The focus is on the overall performance obtained using the Area Under the Curve metric and the feature importance for the top 20 features. Note, all results reported are those obtained on the test set, as opposed to the training set.

As can be expected, the performance is generally higher for shorter time horizons, with the exception of the 72 hour predictive model that outperforms the 24 hour predictive model. It is important to note that all values are significantly above 50% AUC, indicating strong predictive models, and that 1-hour is a very strong model with 89% AUC and a time horizon that is actionable by a future HEMS system.

It is worth noting that whilst the 10 minute time horizon has extremely high predictive performance, 98% AUC, it is not a very accurate representation of reality. The reason for this is that the target variable has been generated by a human, who labels the value in 10 minute increments. As such, it can be expected that any 10 minute period is highly correlated to the previous one. This is likely to be true for a real world setting too, but we expect that the correlation would be smaller in real life. An example of this could be where the human labeller misses someone leaving and returning to the property for a 15 minute period. Given this effect and that the results have been tested on 1 week of data for 1 property, it is important to take care in interpreting the results and not generalising too far.

As compared to the hot water usage predictive models, the occupancy models were built more quickly and smaller emphasis was placed on the feature engineering. As such, a smaller range of features will be observed and for the memory variables only exponentially weight moving averages are used with a half-lives of 2 hours (referred to as “MA” or “moving averages” in the figures).

The key predictive features in the model are historical and current occupancy (auto-regressive term), bathroom humidity, and electrical principal components with memory. Exogenous factors fared well too, which is different to the hot water usage model. At 10 minute and 1-hour time frames, historical and current occupancy hold very strong predictive power, but at the other time horizons this predictive power significantly drops off. Electricity clusters do not add much value.

Time horizon	Predictive performance [AUC]
10 mins	98%
1 hr	89%
4 hrs	78%
24 hrs	63%
72 hrs	69%

Table 8: Occupancy predictive model performance over various time horizons

6.2.1 10-minute time horizon

As can be expected, the most predictive feature at the 10-minute time horizon is current occupancy. This is largely due to the fact that occupant are unlikely to enter and leave a property regularly within a 10 minute time interval, but also may reflect our human bias that humans do behave this way when labelling the data. This auto-correlation largely explains the exceptionally high AUC of 98%. If historical and current occupancy were not included in the model, strong predictive performance would still be achieved, through variables such as time of day and the electrical principal components.

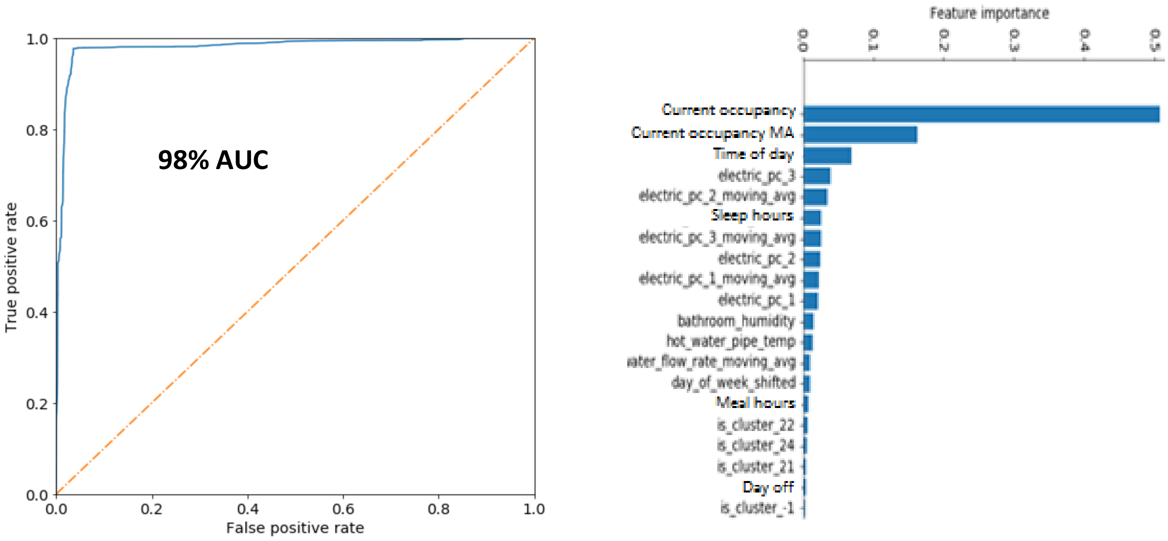

Figure 27: LHS: Receiver Operating Characteristic for 10-min occupancy predictive model. **RHS**: Random Forest ranked feature importance for 20 most powerful features in 10-min occupancy predictive model.

6.2.2 1-hour time horizon

The 1 hour predictive model has a far smaller reliance on current and historical occupancy, but maintains very similar key features to the 10-minute model i.e. time of day and electrical principal components remain key. Performance is still strong, with a AUC of 89%.

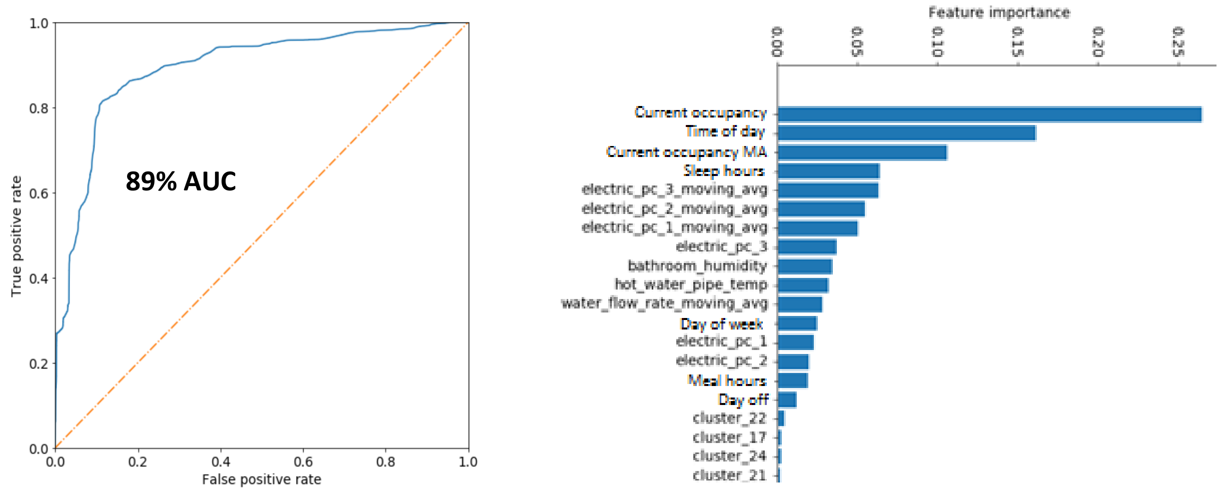

Figure 28: LHS: Receiver Operating Characteristic for 1-hour occupancy predictive model. RHS: Random Forest ranked feature importance for 20 most powerful features in 1-hour occupancy predictive model.

6.2.3 4-hour time horizon

The key features for a 4-hour model remain comparable to the 10-minute and 1-hour models, with the exception of current and historical occupancy, which now hold low predictive power. Predictive performance has reduced, but is still strong at 78%.

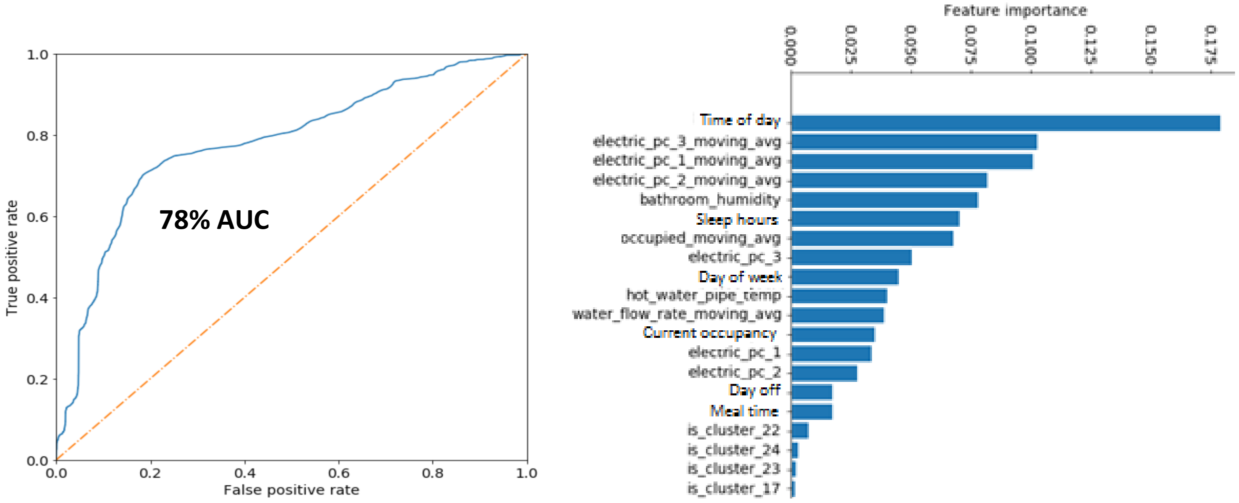

Figure 29: LHS: Receiver Operating Characteristic for 4-hour occupancy predictive model. **RHS**: Random Forest ranked feature importance for 20 most powerful features in 4-hour occupancy predictive model.

6.2.4 24-hour time horizon

The 24-hour time horizon has significantly weaker predictive performance, but still holds value, with a AUC of 63%. The key predictive features relate to humidity, electricity and priors such as time of day and day of week.

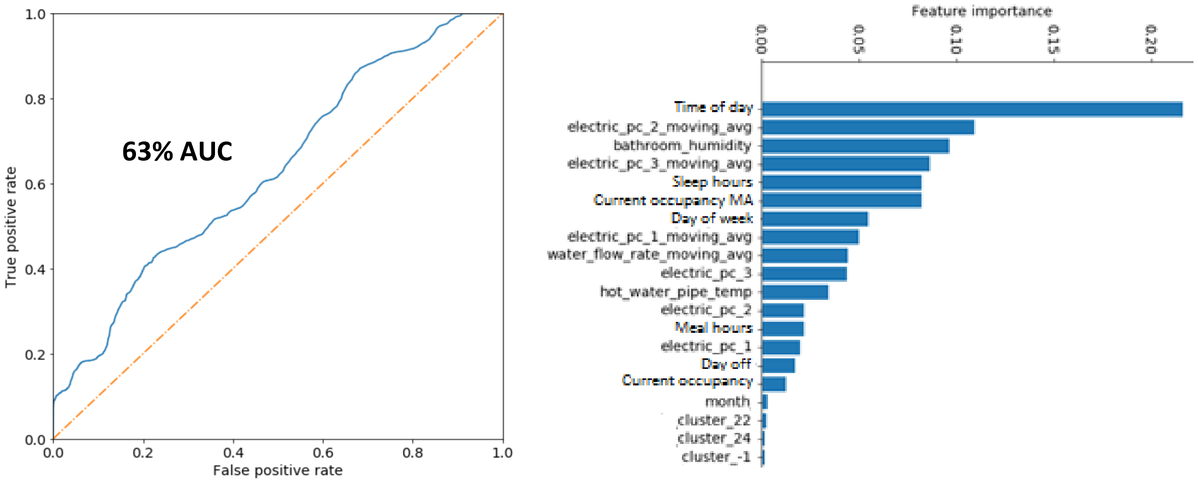

Figure 30: LHS: Receiver Operating Characteristic for 24-hour occupancy predictive model. **RHS**: Random Forest ranked feature importance for 20 most powerful features in 24-hour occupancy predictive model.

6.2.5 72-hour time horizon

The 72-hour model is similar to the 24-hour time horizon model, but with slightly stronger performance, which is likely due to running the test on a slightly different dataset.

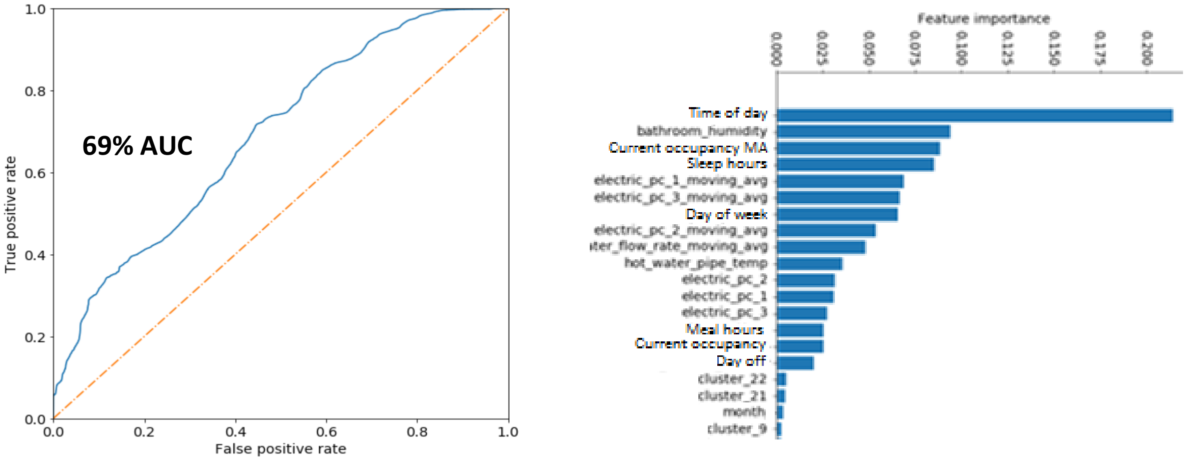

Figure 31: LHS: Receiver Operating Characteristic for 72-hour occupancy predictive model. **RHS**: Random Forest ranked feature importance for 20 most powerful features in 72-hour occupancy predictive model.

7 Potential future work

The focus of the project has been to test the "art of the possible" in predicting future occupancy and hot water usage. The work has not focused heavily on the computational time or cost involved in obtaining the predictive performance, the human intervention and hours that have been required, the need to automate the data pipeline in a production environment, and the commercial value of the solution. All these areas are worth exploring and whilst this report touches on some of them as potential extensions are described, the main focus of the future work chapter is on ways to further improve the predictive models. The chapter is split into 4 sections: datasets; data features; choice of algorithms; and framing the predictive problem.

7.1 Datasets

The project collected approximately 100 Terabytes of data in varying formats, which presented computational challenges from both a volume and variety perspective. Given the various complexities encountered, data was only used for one property, introducing biases in the results reported. It is very possible that the chosen property or period that was analysed is not representative of the broader population leading to misleading performance figures or representation of key data features. Additionally, given that only 30 days of data were used, there was a data sparseness challenge, making it hard to learn certain behaviours and reducing the precision of the results. It is therefore recommendable to extend the analysis to the full timespan of data captured and to all or most of the properties. Additionally, a good part of the HEMS data collected was not used in the analysis including gas consumption, humidity readings for various rooms and temperature readings. These additional datasets could be used as data features to predict

Data quality is another area that could benefit from further work, but is unlikely to deliver significant performance improvements. Firstly, there is an opportunity to do further data quality tests, that may identify new data quality issues which would allow the model to be built more accurately. For instance, tests were not run to check that power readings are not correlated to temperature, which could occur due to faulty devices, or that calls for heat are correlated to gas usage, indicating the setup is correct. Secondly, various assumptions and simplifications were made in processing and syncing the data. For instance:

- **Current readings**: two streams of current readings are provided in the data, at different resolutions. A simple switching algorithm operating at the second level was produced to switch between the two time series favouring the more precise series where possible. A more sophisticated switching algorithm operating at a more granular level could be developed.
- **Time drift**: time drift was corrected for using a linear transformation over the full time period. No tests were run to test whether time drift is linear and the consequent impact.
- **Missing data**: periods of time with poor quality data were removed from the dataset. Over such a short time period, this introduces an element of bias. There are also occasional data gaps spread through the data for HEMS data (humidity and pipe temperature) that require interpolation.
- **Hot water usage**: the target variable of hot water usage is not directly captured in the sensors. As such, it is inferred from water flow and boiler pipe temperature, but is captured as a time period for which hot water is used as opposed to the total volume. Further work could be done to convert this time period to an estimate of volume.

- **Ground truth**: the project intentionally did not aim to capture the occupancy ground truth, to make the experiment more similar to reality, but a future project could consider trying to capture the ground truth so that the true performance of the solution could be measured.

Any future system will need to consider how it handles the data quality issues in a live environment. For instance, if there is a data gap the system will not be able to interpolate between the last data point and the future data point, until it observes that future data point. This detail is not captured in our model. Similarly, if the system sync's its clocks once a month, then the system could not correct for time drift until the month is complete. That said, our assumption is that a future system could load the data, cleanse it and sync it in live, avoiding time drift issues which is key for such a system.

7.2 Data features

Data feature engineering is an excellent way to capture domain expertise in a predictive algorithm and to guide the algorithm to a good solution. Moreover, in this work it was required for data compression. The following are a few areas that could be further explored with the view of improving predictive performance:

- **Appliance disaggregation**: the work explicitly chose to focus on predicting occupancy and hot water usage using the electrical principal components and the electrical state of the property, as measured by 25 clusters. Whilst the approach appears effective, it may be that the approach performs better through the augmentation of appliance level data obtained through appliance disaggregation analysis.
- **Workflow analysis**: 25 property states were identified through the clustering and these were profiled by time of day. An extension here would be to test how the 25 states interact with each other, for instance, does a certain state always follow another. A data exploration exercise of the sort could help inform future data features and it is thought that this may be an excellent way to capture memory in the data, which has proven to significantly increase predictive performance.
- **Additional features**: further feature engineering could be explored including ways to identify when someone enters the house, detecting a motor running or when an LED turns on. Memory also appears to be key, so ways of representing memory and the time span over which these are represented could be explored further.
- **Feature selection**: further work could be conducted to understand the predictive power of various features under different circumstances, informing a future data collection strategy.
- **Data compression**: there is a need for data compression in the electricity data as each property collects around 3 Terabytes of data per month. An approach was proposed that obtained a compression ratio of around 12,000. The approach proposed had the objective of compressing the data as far possible whilst maintaining key information, which was informed by the research team's knowledge of the data and what was felt to be relevant. Future work could investigate if alternative constructs perform better. For instance, does peak binning perform better than peak finding, and was the choice of a 1-second stride and a 2-second time window the right one?

7.3 Choice of algorithm

The predictive model chosen was a Random Forest Classifier, which was chosen due to its strong predictive performance as well as its ability to link this back to the predictive power of the features.

The current implementations have had the minimum leaf size loosely optimised, but there is scope to do further parameter optimisation. Additionally, there is further work that can be done in exploring different algorithms and how learnings from one property is captured in another, there may be a need for a hierarchical representation of learnings.

The random forest classifier was chosen as it has high predictive accuracy in terms of ranking different data points, but the absolute probabilities provided are not always accurate and require calibration. As such, if the algorithm were to be used in a production environment to choose between a few scenarios, it would be important to ensure the probabilities are well calibrated and not affected by up-sampling of certain classes. It is also important to build a mechanism for the various predictive models to be able to effectively interact with one another and be able to incorporate the user's loss function. Another consideration in moving to a production environment would be whether the algorithm should be trained in batch or online mode, which will largely depend on how quickly the algorithm needs to adjust to varying behaviours.

Taking a broader lens to productionising a solution, it is important to think about the danger of converging to a poor solution and how the system owner interacts with the system. A future system is likely to need a way for the human to override the system and provide their feedback, to minimise the impact of a poor solution. Given that the user is likely to have a way of providing feedback, which was not the case in the experimental set-up, it may make more sense to use a reinforcement learning algorithm as opposed to a supervised learning algorithm (as was done for this project). A reinforcement learning algorithm is an agent that interacts with an environment and changes its behaviour based on how the environment responds. This allows for a personalised solution to be obtained based on the customer's interactions. Finally, it is important to have exception handling, which guides the system in the case of unexpected issues i.e. if there is missing data the system may choose to revert to its priors.

7.4 Framing the predictive problem

The objective of the work was to test the "art of the possible" in terms of predicting future patterns of occupancy and residents' needs, in particular relating to heating and hot water. It was necessary to break down the requirement into a clear mathematical formulation, which led to a total of 10 predictive models. Cumulative hot water usage predictive models for five different time horizons, and occupancy predictive models for whether a property would be vacated in exactly the time horizon specified, for five different time horizons. An alternative formulation that could be considered is predicting when the heating is turned on or off, but this may have limited value as it will largely be a reflection of the control algorithm. Additionally, it would be interesting to explore more detailed level occupancy models i.e. which resident is in and in which room. Lastly, the time horizons chosen could be varied.

Finally, it is important to note that in the current formulation 5 fixed time windows are used and the quantity (occupancy or hot water usage) is predicted. An alternative would be to fix the quantity and to predict the time i.e. how long till the hot water is used for 10 minutes. The two formulations will have different performance as the distributions around these two target variables will be different and it is worth testing how the results compare.