Mining Household Characteristics With Domestic Electricity Consumption Data

Sam Stern (s1134468)

THE UNIVERSITY
of EDINBURGH

April 2, 2015

Abstract

Smart meters are being rolled out to millions of households across the UK to monitor and relay electricity and gas consumption data back to energy suppliers. Whilst the smart meters can help consumers control their energy consumption and improve transparency in the energy providers' billing practices, concerns about what can be inferred from the data are being raised. This project justifies these concerns by showing that it is possible to automatically predict household characteristics from smart meter data. Focusing on identifying the presence or absence of children in a household and its socio-economic status, supervised learning techniques are used to construct classifiers and evaluate them using data from the Household Electricity Survey. We show that it is possible to extract features that can distinguish between different household classes. These features are then used to construct classifiers, able to infer whether there are children present in a household has children with an accuracy of 83.6%, and Matthews correlation coefficient (MCC) of 0.6 and a household's Ipsos MORI social grade with an accuracy of 53% and MCC of 0.41.

Contents

Chapter 1

Introduction

1.1 Introduction

Amidst international pressure on countries to reduce their carbon footprints [1, 2] and the British public's increasing frustration over poorly explained yet continuously rising residential energy bills [1], the UK Government is currently executing an £11bn plan to distribute 53 million smart meters to more than 30 million households and small businesses across the country by 2020. Smart meters measure a household's energy consumption and wirelessly transmit the readings to suppliers in 30-minute intervals. The initiative is expected to help households control their gas and electricity usage by communicating how much is being used and the associated costs, often on a per-appliance basis, to in-home display monitors. Smart meters should also increase transparency in domestic energy billing by eliminating the need for monthly meter readings and usage estimations. Instead, energy companies receive smart meter-generated accountings of their customers' real consumption, from which they should be able to invoice more accurately.

While there has generally been strong support for the smart meter campaign [2], there has also been resistance, with fears that energy companies will use the information as an opportunity to raise customers' bills and increase their own earnings [3, 4]. These concerns are not completely unfounded. Already, data management companies, such as Buckinghamshire-based SAS, are promoting their ability to extrapolate information from, and exploit, smart meter output, offering energy suppliers "vital insights...for customer retention, targeted marketing and increased profitability...based on actual [customer] behaviour"[5].

Even well-intentioned programmes run the risk of crossing privacy lines. SHIMMER[1][6] is an energy management scheme for smart meter-equipped households in the country's lowest socio-economic classes that helps them save money, in part by piggybacking an on-line interface to its open system network onto the home's communication hub to help "control and automate appliances, central heating and household finances" and over time, act as a "portal" to on-line banking, lender, and home improvement providers.

In addition to financially-motivated breach-of-privacy issues, such as those posed by SAS, or worries over Big Brother-type government involvement, as in the case of SHIMMER, other threats related to a steady stream of smart meter data include burglaries, cyber-attacks and other crimes of opportunity.

[1]Smart Homes Integrating Meters Money Energy Research

Addressing these issues, the Government have incorporated layers of data access, sharing and usage regulations and restrictions as well as numerous wireless communications protocols into the national smart meter programme framework [7]. These controls notwithstanding, privacy questions do remain, with one of the most fundamental being: Just how much information about a household could realistically be inferred from raw energy consumption data alone?

In looking for an answer, this project explores whether (and to what extent) it is possible to construct a model that accurately predicts detailed personal information about a household based on its energy consumption readings and, if so, if the results would be reliable.

Using energy consumption information collected by the Household Electricity Survey (HES), a Government-sponsored national survey of domestic electricity use collected over the period 2010 to 2011, as the dataset, classification models are created to predict two properties of households: 1) The presence (or absence) of children and 2) the Ipsos MORI social grade of the chief income earner. These properties have been specifically chosen because, of all the information gathered by the HES survey, they would logically be of interest to someone who might wish to intrude on a household.

The project has 3 main components:

1. Clean the HES data and create a database that stores each households' energy-use information and any other relevant data;
2. Extract useful features from the data that can be used as inputs to a classification model;
3. Build models to infer household properties using supervised learning methods and evaluate the results.

1.2 Smart Meters

Smart meters are digital energy usage measurement devices with integrated wireless two-way communication components that allow them to be interactive, interoperable and remotely readable and programmable [4]. As applied to the UK domestic energy sector, however, the term 'Smart Meters' refers to a bundle of equipment and technologies that not only tracks electricity and/or gas consumption, but makes use of readings, billing, tariffs and other related information available to both households and their energy suppliers[8].

A dedicated wide area network (WAN) sits at the heart of Britain's smart meter grid, transferring data between energy suppliers and their customers. Households are fitted with a home area network (HAN) communications hub to which are connected a HAN-WAN interface, the dwellings' smart gas and/or electricity meters, and at the customer's discretion, a portable in-home display monitor and selected appliances.

At the other end of the WAN are the energy suppliers, who are responsible for organising the supply, installation and servicing of the smart meters, monitors and hubs in customers' homes[8].

The type of information collected on smart meters includes[4]

- real-time and historical energy consumption, both on a mains and individual appliance basis
- current and historical payment information and banking details
- energy tariffs and types of metering (e.g., pre-paying or direct debit)
- customer profiles

Smart meters are not novel to the UK. EU Countries such as Italy, Sweden, Finland, Switzerland and Germany have already begun implementing their own roll-outs, and smart meters have been in use for several years in many US states and Canadian provinces. One of the reoccur issues is the granularity of data provided to the suppliers. In the UK, it was determined that since the suppliers purchase energy in 30 minute blocks, that this would also be sufficiently detailed for billing customers[7, 9].

1.3 Related Work

Particularly in recent years, an increasing number of studies have applied machine learning and data mining techniques to model and analyse domestic electricity consumption. This field of research is of particular interest to energy providers, as understanding who their clients are and how and when they use energy lets the providers optimise their resources (providing more power during peak times and less during periods of low demand) and create and, as alluded to earlier, market products to specific client groups.

The research done using household energy data can be broadly separated into two categories whereby either: 1) only consumption data is analysed to categorise households or 2) the data is related to additional information about the household. The first approach imposes fewer requirements on the data itself and has therefore often been used in unsupervised tasks [10]. For example, Chicco, in a study analysing electrical load pattern data, gives an overview of the clustering techniques used to establish suitable client groups [11]. Cao et al. also grouped consumers using electricity load profiles, however they focused on finding households with the same peak usage [12].

Another popular area of research is NILM (non-intrusive load monitoring), which involves taking aggregated energy consumption data from households and disaggregating it to find the load used by constituent appliances. Kolter and Jaakkola employed factorial hidden Markov models (FHMMs) to disaggregate energy readings, achieving more than 90% precision on a synthetic data set [13]. Similarly, a study performed by Lisovich et al. used NILM to determine when people were present in a household, the appliances they were using (and when), and their sleep/wake cycles. This was done by looking at a dataset of dwellings that had energy readings taken at either 1 or 15-second intervals for between 3 and 7 days. Compared to the dataset used in this report, however, the households in the Lisovich et al. study were more homogeneous, particularly with respect to appliance types (e.g., none of the Lisovich et al. households used electric showers or water heaters) [14].

McLoughlin et al. explored the correlation between electricity consumption data and household characteristics using a dataset of smart meter readings taken from 4,232 Irish households. They also investigated methods for clustering households based on energy use. Beckel et al., with the same dataset as McLoughlin et al., used supervised learning methods to classify household attributes. Their research involved predicting a set of characteristics describing the inhabitants, such as the age of the chief income earner, presence/absence of children and socio-economic status. They also sought to identify properties of the dwelling itself, including the number of appliances and bedrooms, and the types of cooking facilities [10].

In contrast to the Beckel et al. study, the work being presented here, while tackling some similar issues, considers a different set of classifiers (random forest, logistic regression and Knn) as well as a new class of features taken from the time-frequency transform of the data. Another distinction is that, to boost performance, Beckel et al. considered models that relied on *a priori* knowledge beyond that available from electricity metering alone, whereas no other prior knowledge outside of the data recorded by the smart meters is assumed in this report. Finally, while Beckel et al. used a dataset of Irish households, the HES dataset contains only English residences. This distinction is important because smart meter implementation practices vary widely from country to country, meaning the results obtained by Beckel et al. are not necessarily transferable to Britain [4, 15].

1.4 Project Description

The goal of this project was to address the privacy concerns that have been raised regarding smart meter data. More specifically, the project looks at whether it is possible to create a probabilistic model that can automatically predict intimate information about a household given only features that can be extracted from the data available from smart meters. Supervised learning methods were used to predict the socio-economic status of a household and whether there are children present. In this paper, we argued that, while one can construct a model that infers household characteristics, numerous factors contribute to the way in which a household uses energy.

The remainder of the report is structured as follows: First, the HES dataset, which contains labeled electricity consumption data from 250 households, is introduced. The steps performed to pre-process the data are then presented, along with challenges encountered and issues that needed to be accounted for (as pertaining to the dataset). Next, the methods used to extract features from the HES dataset are outlined, and feature selection methods are discussed. The classification models are then introduced and optimized, before discussing their performance on unseen household data. Finally, the results are compared to those obtained by similar studies and the original question regarding privacy invasion is discussed in relation to the results.

Chapter 2

Data

2.1 Overview of the HES Dataset

The data used in this project comes from The Household Electricity Survey (HES), a UK government-sponsored study of residential energy usage jointly commissioned by the Department for Environmental, Food and Rural Affairs (Defra), the Department of Energy and Climate Change (DECC) and the Energy Savings Trust. HES tracked the electrical power sources and demands a variety of owner-occupied homes in England over the period May 2010 to July 2011 [16, 17, 18]. The study sought to identify, catalogue and analyse the range, quantity and energy requirements of appliances found in 'typical' British homes, with the underlying aim being to better understand households' frequency and patterns of usage, and to collect any 'user habit' and/or other socio-economic data that might emerge [18, 19]. Information would be leveraged in numerous ways and for a variety of purposes, not the least of which would be as an aid in developing energy policy (both at the consumer and energy provider levels), and help justify the cost of the smart meter roll out.

The HES study monitored 250 households, of which 26 were observed for one year, and the remaining 224 for roughly one month. Although all the participating households were located in England, they did not share the same demographic or geographic profiles. This was reflected in the wide spectrum of number and ages of appliances. Whereas one household, for example, registered just 13 appliances, another had 85. There was a forty-one year old freezer, several brand new televisions and a broad assortment of ages and types of devices in between. When aggregated (as outlined in Section 2.2), the result could be considered an estimate of an average mains reading.

Smart meters record the total energy being used in a given interval, whereas when discussing individual appliances, it is common to talk about the energy used per unit time (i.e., power). For example, an average kettle might use 3kW of power. If the kettle is running for 2 minutes then the energy used would be 6kWm (kilo Watt minutes). A potential issue arises when we consider another appliance that uses less power but for a longer period of time. A hairdryer, for instance, uses 1kW of power. If a hairdryer was being used for 6 minutes, then the total energy used would also be 6kWm. Although the hairdryer uses less power than the kettle, the smart meter reading would record the same number (6kWm). The smart meters used in the HES study took readings either every 2

minutes or every 10 minutes in units of deciwatt hours (dWh or 0.1Wh). This is was measure of the total energy that the homes' appliances consumed since the last reading. As it is conventional to describe the energy used in terms of kilowatt hours (kWh), the readings are each divided by 10,000.

In addition to the data collected on appliance types and the meter readings, participating households also kept diaries of how they used their mains appliances, and provided supplemental information about the households' constellations, such as: the number of occupants, employment status, Ipsos MORI social-grade and whether there were children present.

2.2 Extracting the Data and Pre-Processing

As explained in Section 2.1, electricity readings from individual appliances and sockets were taken for each household. This was in contrast to the total-energy-consumed figures needed for this project. In organizing its data, the designers of the HES study assigned values to the 250 possible appliances that the designers of the study that they expected a household to have. Appliances that were not present were designated a 0. The resulting raw data was held in large csv files. Since no household had all 250 potential appliances, there were a significant number of redundant entries. To use the data to perform data mining for this project, numerous pre-processing steps needed to be performed. This was accomplished by writing and implementing python scripts with embedded SQL.

The first step in pre-processing the data was to create a MySQL database and import the appliance readings into a table. Cambridge Architectural Research Ltd (CAR) [20] , an architectural consultancy, provided additional files that mapped which appliances needed to be aggregated for each household to arrive at an estimated mains reading, as this was often not simply the sum of all appliances' readings. A table was therefore created for every household where each row contained the aggregated electricity measurements for a given date and time.

Another consideration was the number of participating households. 250 is a relatively small sample size for a machine learning task, and would be likely to over-fit to the sample population. To help account for this, the 26 households that were monitored for an entire year were split into 12 instances that could be treated as separate entities. This generated an additional 281 household instances. While it did not create a more diverse group, it did add more instances to train and validate, as well as with which to test a classifier. To avoid over-fitting the classification models to the data, all instances from any one of the 26 (split) households were either in the training or test set, but never in both.

The inconsistency in measurement intervals mentioned earlier in this chapter also had to rationalised. While some households reported how much energy they used every 10 minutes, others were measured in 2-minute intervals. To create consistency in the data, for the '2-minute households', every five intervals were summed so that all the households had 10 minute granularity. This step was important because some of the consumption features would have been affected by differences in measurement intervals.

The last stage in pre-processing the HES data was to ensure that each instance was of the same duration. This included making certain that every day occurred the same number of times throughout the study (e.g., four Mondays,

four Tuesdays, etc.) for every household. It also had to maintain its sequential order within the week (i.e., Mondays had to always follow Sundays and come before Tuesdays.) There are several reasons for this step. When visualising the data, temporal structure was observed both intraday and intraweek. In addition to an obvious daily pattern (more energy being used during the day than at night), there was repetition over weekly cycles, where it was possible to distinguish some days of the week from others, as a pattern emerged with a 7-day period.

Ensuring each household had an integer number of weeks meant that no single day of the week would affect the total energy used more than the others. For example, the data that more energy was used on Sundays than on any other day of the week for most of the HES households. So, if one household had three occurrences of a Sunday during the monitoring period, but another had four occurrences, then when computing features, such as average daily energy, the values obtained would be different for the two households simply because of the extra Sunday. Because most of the households were recorded for roughly one month, it was decided to ensure that the data for each household was 28 days long (4 weeks). Again, it was important to make certain that the days of the week remained in order. Some features (such as computing the Fourier transform) expect the data to be in sequential order. Finally, the data was arranged to start on the same day of the week for each household, as this made it far more convenient to extract specific parts of the data.

Based on this logic, the following steps were used to pre-process the HES data:

1. Ensure that each household had an integer number of days by topping and tailing the data.

2. Find the mode day of the week that the data started from (this was found to be Sunday).

3. For the households that do not begin on a Sunday, chop the top few days so that the data does begin on a Sunday.

4. If the household's data was now less that 28 days, append days-to-the-end until it is of the correct length (If possible, used the days that were chopped off in the previous step; otherwise, reused a day's worth of readings).

Figure 2.1 gives a visual example of data that has been made to be of uniform length. As the readings start on a Thursday (Day 5), the first three days are chopped off the top. Since the data is now less than the required number of days, days are either reused or, if possible, taken from the days that have been chopped from the top[1].

[1]It was noted that this method creates a bias in the features. For example, when computing the average energy used on a Monday, if a household only had three unique instances of a Monday and one instance had been reused, then this would have affected the feature.

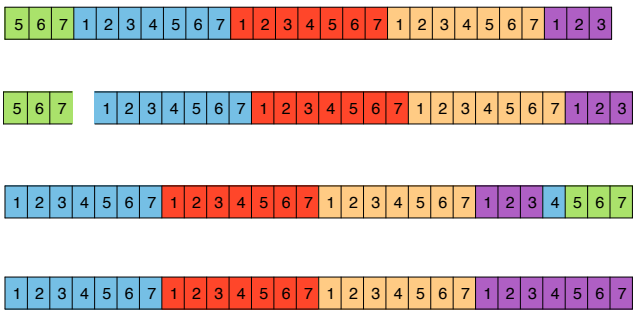


Figure 2.1

2.3 Household Classes

Each household that participated in the HES study completed a survey including questions about the dwelling they occupied (such as the year it was built), the household itself (such as the number of occupants) as well as their attitudes toward climate change and energy consumption. The answers to these questions are used as labels for the households in performing supervised learning for this paper.

Social Grade	Description	Sample Size	% Sample	% UK Population
A	High managerial, administrative or professional	33	6.4	4
B	Intermediate managerial, administrative or professional	95	18.3	23
C1	Supervisory, clerical and junior managerial, administrative or professional	197	38.0	29
C2	Skilled manual workers	128	24.7	21
D	Semi and unskilled manual workers	34	6.6	15
E	State pensioners, casual or lowest grade workers, unemployed with state benefits only	32	6.2	8

Table 2.1

Tables 2.1 and 2.2 show the sample sizes for each class of the two classification problems being considered in this project. The distribution of households over each of the classes in our sample is similar to the true distribution, which means that the empirical prior probability of each class is a reasonable estimate of the true prior probability. However, there is a significant imbalance in the classes, especially in the socio-economic classes. This will result in a bias in the classification models that will need to be considered when evaluating them.

Class	Sample Size	% Sample	% UK Population
Children	187	36	39
No Children	332	64	61

Table 2.2

2.4 Discussion

After the data had been extracted from the csv files, pre-processed and imported into MATLAB, plots of the data were made to visually gain insight into how households had used energy and to increase domain knowledge. Figures 2.2 and 2.3 are examples of how some of the households consumed energy. Both figures show the data gathered from the same households, but over different time periods. Studying these plots gives valuable insight, which is used later to aid in feature extraction as well as to ensure that the data appears reasonable.

In Figure 2.2, the first thing we notice is that the consumption is not smooth. There are sharp peaks that vary in height, which can be used to make assumptions about which appliances are being used. For example, many of the peaks are around 0.1kWh, which is roughly the amount of power used by a kettle (approximately 2kW for 3 minutes). The next thing to note is that there is an obvious underlying daily repetition. The household tends to use more electricity at night than it does during the day time. Finally, it can be seen that the energy consumption on weekends is slightly different than that of week days. There are short periods of abnormally high electricity on Saturdays and Sundays which are observed less frequently during the week. To see this, note that both Figures 2.2 and 2.3 start on a Sunday, and that each 'wavelet' is one day long.

It was because of these observations that the data was made to be four weeks long. That meant that each day of the week appears exactly 4 times for each household means that features such as the total energy used are not influenced by which days of the week are present.

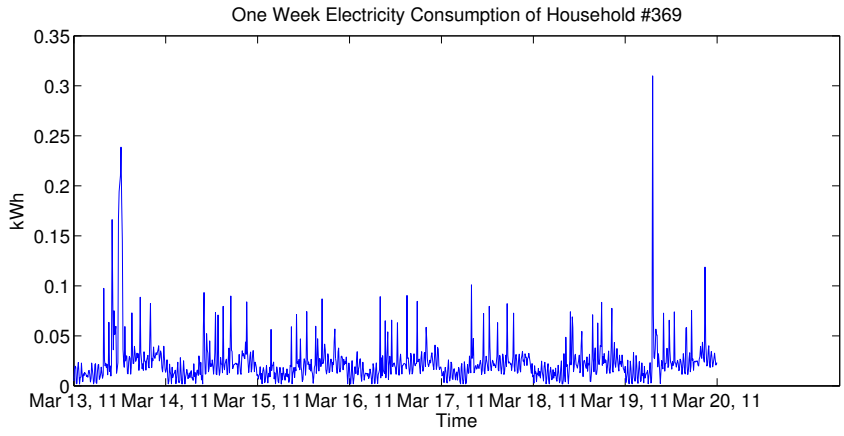


Figure 2.2: The electricity use of household # 369 over a week after being pre-processed

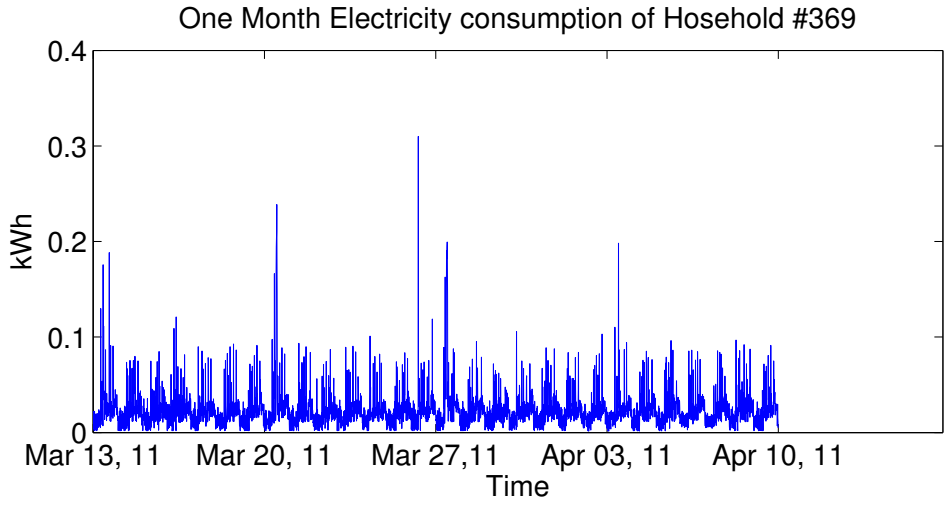


Figure 2.3: The electricity use of Household # 369 over a four week period after being pre-processed

2.5 Issues

As with any data mining project, issues arose which needed to be recognised and dealt with. Almost invariably related to the source data, some of the challenges were simply the result of environmental influences, while others reflected the design and implementation of the source study. Those most relevant to this project are outlined below; more can be found in the CAR report [19].

The first hurdle came about due to the relatively low number of participating households. Comparing the HES Study's 250 households to, for example, the 4,232 that took part in Ireland's CER (Commission for Energy Regulation) study of household electricity consumption[21] (used by Beckel et al. and McLoughlin e al. [22, 23, 10, 24]), it is less likely that the UK results generalised as well as the Irish ones, particularly for the multi-class classification problem (where there were as few as 32 households per class).

Moreover, only English homes were included in the UK study (i.e., Scotland, Wales and Northern Ireland were not represented) and all of the houses were owner-occupied. While 84% of the British population does live in England, only 64% of English residences are owner-occupied [25]. Suffice to say, the subset of participants considered in the HES study was not fully representative of the UK as a whole. It is important to remember, though, that the aim of this project was to determine *whether it is possible* to infer a household's properties from its electrical power consumption, not to build a classifier that could be used to infer household properties from smart meter data. This is an important distinction, as this was a proof-of-concept project, intended to look at whether information about a household is contained in its energy-use patterns, as opposed to an attempt to build a commercial product. As such, the quality of the sample population (or lack thereof) was not necessarily detrimental to the aim or outcome.

The HES study involved recording individual appliances, rather than mains readings, so the total electricity consumed by each household could not be given with certainty (as there was no way to confirm whether all the appliances and sockets in a household were being recorded). The effect of this complication can

be seen in Figure 2.4 where Household #75 is always using at least some energy while Household #121's consumption drops at times to 0. Household #75 offers a better estimate of what should be expected from a realistic household, as it is reasonable to assume that there will always be a minimum amount of electricity used by a home (since no appliance is 100% efficient and will usually leak at least some electricity).

A related issue had to do with the quality of the readings taken, particularly with respect to the influence that the individual appliances chosen to be included in the study had on each participating household's total electricity demand profile. Where the readings depicted in Figure 2.3 appear to credibly demonstrate the characteristics of a typical home's energy consumption, that home, for the purposes of this project, one of the 'better' households, and does not reflect the quality of the samples across the board. Figure 2.4, on the other hand, shows examples of homes that did not follow the same kind of trend. Looking at their reconstructed mains readings, these households either did not have well-defined periodicity or they used significantly more (or less) energy than one might expect of a household. For samples such as these, the task then became finding a means of determining whether the discrepancies were reasonable differences that could be attributed to natural variations between households, or whether they were the result of poorly executed data collection.

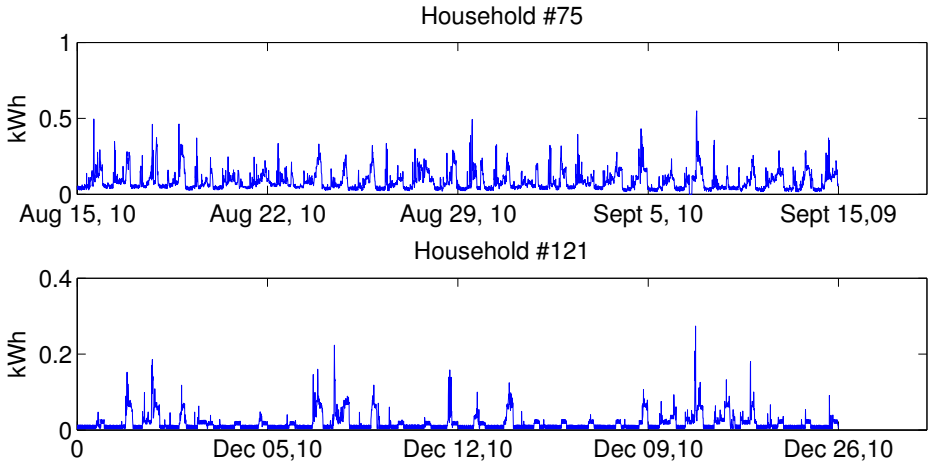


Figure 2.4: The electricity consumption of two households that do not show the same pattern of consumption as other households

Table 2.3, taken from a Centre for Sustainable Energy [26] report, shows the amount of power various appliances draw. Comparing these values to the HES readings, it is reasonable to expect a household to use anywhere from 50W to upwards of 15,000W. However, some of the most electricity-expensive appliances, such as electric cookers, electric showers, electric heaters and tumble dryers (which have large affects on household consumption) were not present in all sample households. Electric water heaters, for example, were present in only 38 of the 250 households. This was important to keep in mind when evaluating the models, as such latent variables would have had an impact on a household's electricity consumption.

Table 2.3: **Energy used by various household appliances**

Appliance	**Rating**		**Appliance**	**Rating**
Immersion heater	3,000W		Fridge	40-120W
Electric fire	2,000-3,000W		Fridge-freezer	200-400W
Oil-filled radiator	1,500-2,500W		Freezer	150W
Electric shower	7,000-10,500W		Electric mower	500-1,500W
Dishwasher	1,050-1,500W		Electric drill	900-1,000W
Washing machine	1,200-3,000W		Hairdryer	1,000W
Tumble dryer	2,000-4,000W		Heating blanket	130-200W
Toaster	800-1,500W		Games console	45-190W
Kettle	2,200-3,000W		Laptop	20-50W
Microwave	600-1,500W		Desktop computer	80-150W
Oven	2,000-2,200W		Tablet (charge)	10W
Grill/hob	1,000-2,000W		Broadband router	7-10W
LCD TV	125-200W		Smart phone (charge)	2.5-5W

In examining the data closely, a small number of instances could be noted where the electricity usage for selected households showed consumption levels either flat-lining or completely disappearing for one or more days. To compensate for such anomalies, the usage patterns of each affected household before and after the aberration were analysed and compared to the event. In most cases, it was possible to visually determine whether the incongruity was more likely related to a residence being unoccupied (i.e., the family were on holiday), or if it was due to some technical issue(s) linked to the energy readings themselves (i.e., the meters were presenting erroneous data). Because these sorts of missing data would affect the resulting models trained on that data, the decision was taken to discard households with a consumption reading of 0kWh when that reading represented a statistically significant proportion of the total time for which the household was being observed. If a reading for a single day appeared completely out of keeping with all the other readings for that household, then that day's data was discarded and replaced by an equivalent day from another trial week.

Finally, one factor that had less impact on the data than anticipated was seasonal weather fluctuation. Lower temperatures and shorter periods of sunlight during colder months have been shown in studies to precipitate higher energy usage [27]. Although CAR was able to provide a document outlining which appliances needed to have their readings adjusted to account for seasonal factors, this was not accompanied by any justification for why the specific values had been chosen and did not appear to influence the data significantly. Since most households were recorded in the colder months between November 2010 and April 2011, and those that were measured for a year did not appear to significantly change their consumption in the warmer months, seasonal adjustments were disregarded.

Chapter 3

Feature Exploration and Extraction

3.1 Types of Features

When data mining in time series, considering each point in time sequentially as a feature is not usually sufficient. Aside from ignoring the high dimensionality of the data, it does not account for the correlation between consecutive values [28]. It is therefore beneficial to transform and aggregate the data in such a way as to reduce the dimensionality as well as capture differences in consumption patterns between classes.

According to Beckel et al. [23], features of interest when classifying households based on energy consumption would include: consumption figures, ratios, temporal properties, and statistical properties. Consumption figures represent the average, maximum and minimum amount of energy used over a defined period of time. Ratios are features that calculate the fraction between the consumption figures; they capture relevant patterns that occur across points in time. Temporal features express the first or last occurrence of some event, the time at which the daily maximum or minimum energy demand takes place, or any periodicity within the household's electricity consumption. Finally, statistical properties, such as variances or correlations, give insight into the consumption curve and demonstrate how a household's activity may fluctuate throughout a predetermined time period.

3.2 Non-Linear Transformation

Numerous statistical methods presume that the input follows a normal distribution. With this in mind, the HES data was visualized and compared against a normal quantile plot to find suitable non-linear transformations [29] [30]. Figure 3.1 shows the normal quantile plot of the average standard deviation of a household on Mondays (left) and the logarithm of this feature (right). The linearity of the sample quantiles of the features (x-axis) versus the theoretical quantiles of a normal distribution (y-axis) implies that the transformed features are (roughly) normally distributed. These transformations are important for classifiers, such as Knn which rely on the distance between samples based on their features.

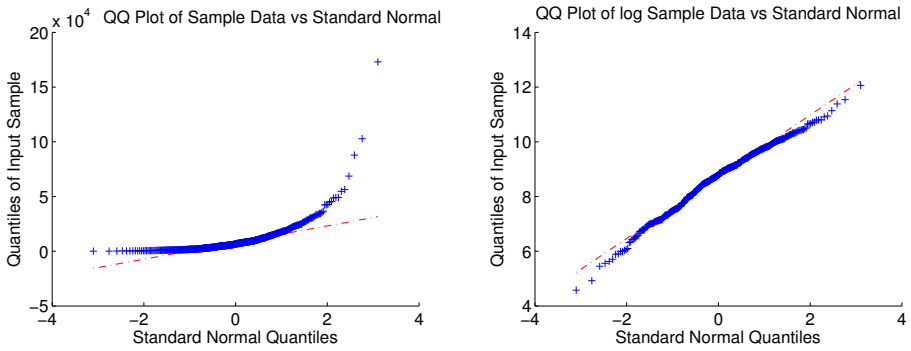


Figure 3.1: The Q-Q plot compares two probability distributions of the data to that of a normal distribution

3.3 Creating Features

One method of extracting features is to compute as many different combinations and aggregations of the data as possible, compare them all and then choose those that best discriminate between classes. Applying this to households, in addition to considering the energy used over a month, data could be further split into weeks, days and even hours. Consumption figures and statistical properties would then be calculable for each of these intervals. While this method does provide more coverage, and thus a greater probability of finding the optimal features to discriminate between classes, it ignores any domain knowledge that we might already have. As such, it is potentially wasteful of the limited resources available for the project at hand.

Focusing on efficiency, instead of computing features in an ad hoc manner, they were created as follows: 1) Assumptions were made regarding the distinction between classes (e.g., households with children use more energy overall). 2) Features were created to capture this distinction (e.g., the average energy over a 4-week period). 3) Tests were performed to evaluate the validity of the assumption (e.g., visualising the feature). The tests varied in thoroughness as it was sometimes obvious from visualisation alone that they discriminated between classes. At other times, more sophisticated methods were needed, as described in Section 3.5.

The remainder of this chapter describes features that were created from the energy reading data and justifies why it was assumed that they would discriminate between classes. These are shown in the form of a box-and-wisker plot. The y-labels were omitted purposely cod the consumption figures that were log-transformed as this makes them unitless. The results of computing these features are then evaluated. Both classification problems (socio-economic classification and child classification) were considered when choosing features to evaluate.

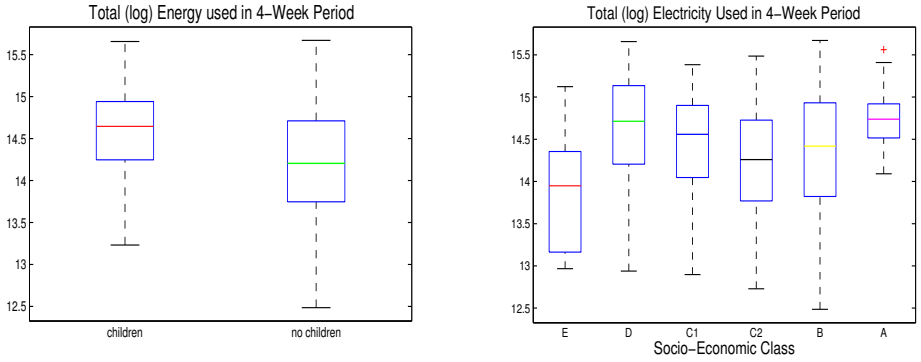
Total Electricity

In visualising the HES data, it became evident that there were large disparities in energy usage among households. While some had an average smart meter reading of 109Wh, others averaged as little as 40Wh[1]. One household was recorded to

[1]The smart meters actually measured in units of 0.1Watt hours

have consumed as much as 3.25kWh in 10 minutes (3,250Wh); another never used more than 0.198kWh over the same time interval. To determine whether these discrepancies could be attributed to different classes, the first feature that was explored was the total energy consumed within a given period of time. This was chosen to be 28 days (for reasons outlined in Section 2.2).

Building a classifier using the total electricity as input assumes that some classes use more energy than others. This can be justified as correlations between a household's disposable income and the amount of energy it uses have been observed in previous studies [31].



(a) Total electricity used by households in a 28 day period, grouped by whether the household has chidren or not.

(b) Total electricity used by households in a 28 day period, grouped by the IPSOS social grade of the household

Figure 3.2: Total electricity used by households over a 28-day period grouped by the presence of children (left) and socio-economic status (right).

Looking at Figure 3.2, there appears to be a distinction between classes in total electricity consumption. The left hand plot, which compares households with children against those without, indicates that those with tend to use more energy. The right-hand plot, which compares total electricity according to social grade, shows that the highest socio-economic households use more energy than those of the lowest social grade. It does not, however, discriminate well between intermediate classes.

Average Daily Usage

Having established that some classes of households do indeed use more energy than others, the average energy used by each household on each day of the week can be computed to determine the underlying cause. This sort of feature explores not just whether some classes use more electricity than others, but if the electricity consumed is dependent on the day of the week. This was computed by taking the total energy used on each day, grouping similar days together (based on which day of the week they are), and then taking the average of each group.

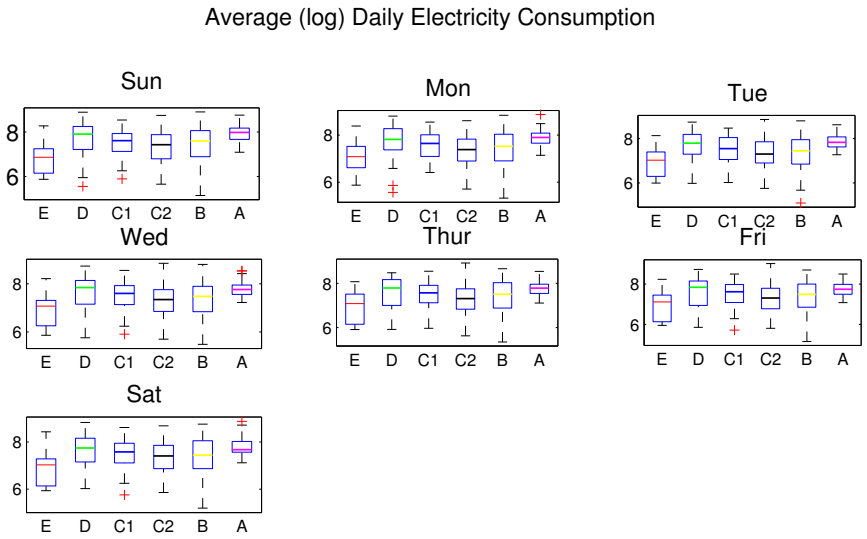


Figure 3.3: The total energy used per day, averaged over a four-week period, and grouped according to the presence of children. 'C' refers to households with children and 'NC' refers to those without children.

Figure 3.4

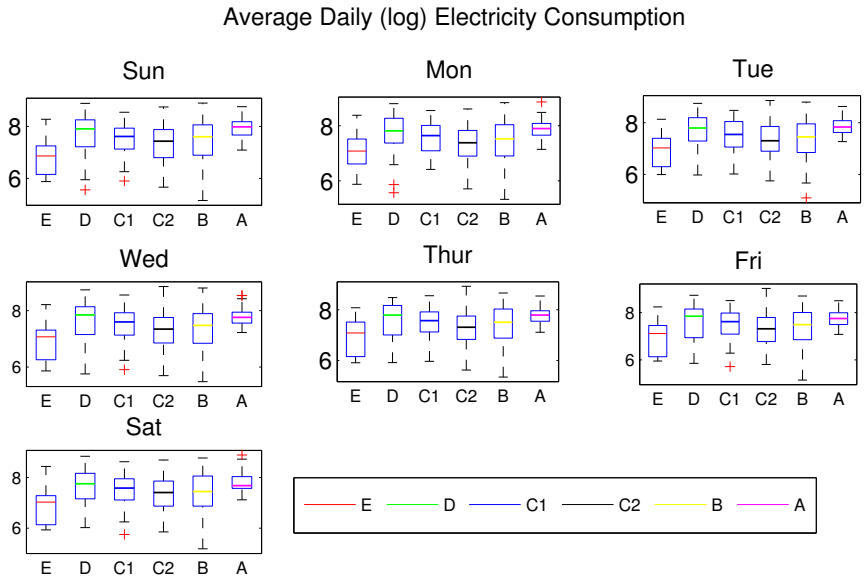


Figure 3.5: The total energy used per day, averaged over a four-week period, and grouped according to their Ipsos mori social grade

Figure 3.4 further illustrates that households with children use more power than those without. It does not, however, provide any additional insight as to when, how or why this is the case. Households with children tend to use more electricity per day, irrespective of what day of the week it is.

Similarly, Figure 3.5, which compares the average daily usage of among socio-economic groups, does not offer any new insights into the differences between

classes. There is no particular day where differences in electricity consumption between classes are more visible.

It should be noted, that due to the way the data was pre-processed in Section 2.2, there is an inherent bias in features that rely on average consumption over the four-week period. If a day had to be reused in order to ensure 28 days worth of data, then it is counted twice when taking the average.

Average Part-Of-Day (APOD)

Going further, different classes potentially use more or less energy at different times of the day. For example, lower socio-economic households might use more energy during the day than those of medium or high socio-economic status. This is because the conditional probability of unemployment given lower socio-economic status is higher than that of unemployment given higher socio-economic status [32]. Similarly, it would not be unreasonable to assume that the consumption gap between households with and without children might shrink when the children are at school and widen when they are at home. APOD (Average Part-Of-Day) features are computed by grouping the smart meter readings according day-of-week (as in the previously computed feature), as well as by part-of-day. The resulting averages of each group is then taken to be a feature. As detailed below, it makes sense to look at days as the sum of four parts, resulting in 28 new features.

Most schools days in England begin at 9:00 and finish between 15:00 and 16:00 [33]. Combining this information with an assumption that as children go to bed, the activity of the other members of a household will decrease and therefore electricity consumption will drop, it is worthwhile to split each day into the following groups.

1. Morning (6:00-9:00): The time when members of the household wake up and prepare themselves for work, school etc.

2. Daytime (9:00-15:00): The time children are at school.

3. Evening (15:00-22:00): When a household can be presumed to be most active

4. Night (22:00-6:00): Depending on the type of household, the time when people might be more or less active. For example, couples without children might stay up later.

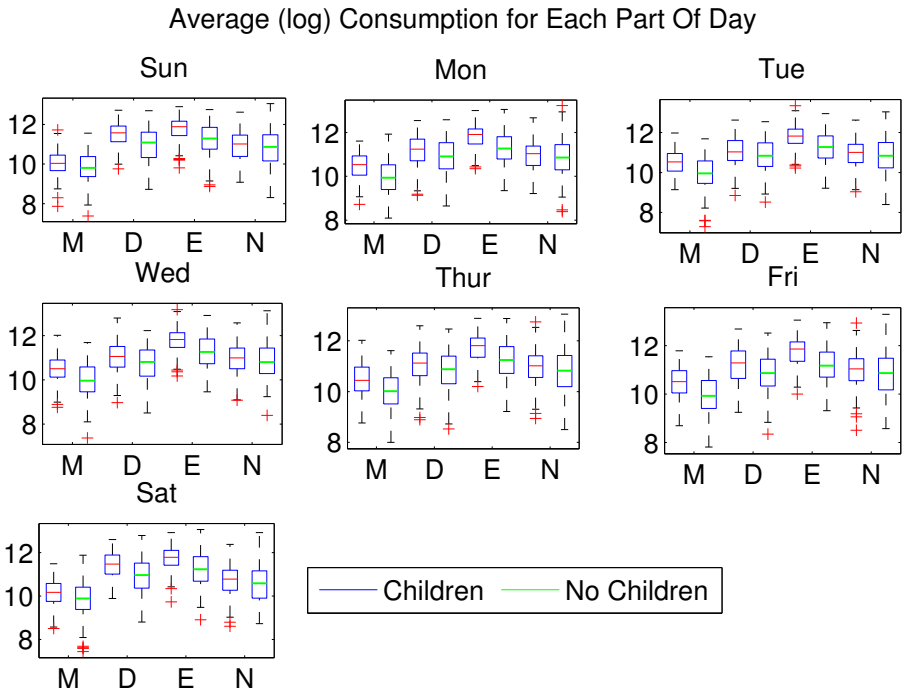


Figure 3.6

The data portrayed in Figure 3.6 indicates that energy use patterns are indeed different for households with versus without children. Much of the difference can be attributed to household activity in the Evenings (15:00-22:00). It can also be seen that on weekday Daytime (9:00-15:00, Monday-Friday), the two classes use similar amounts of electricity. However on Saturdays and Sundays, the gap widens and those with children tend to use more than those without.

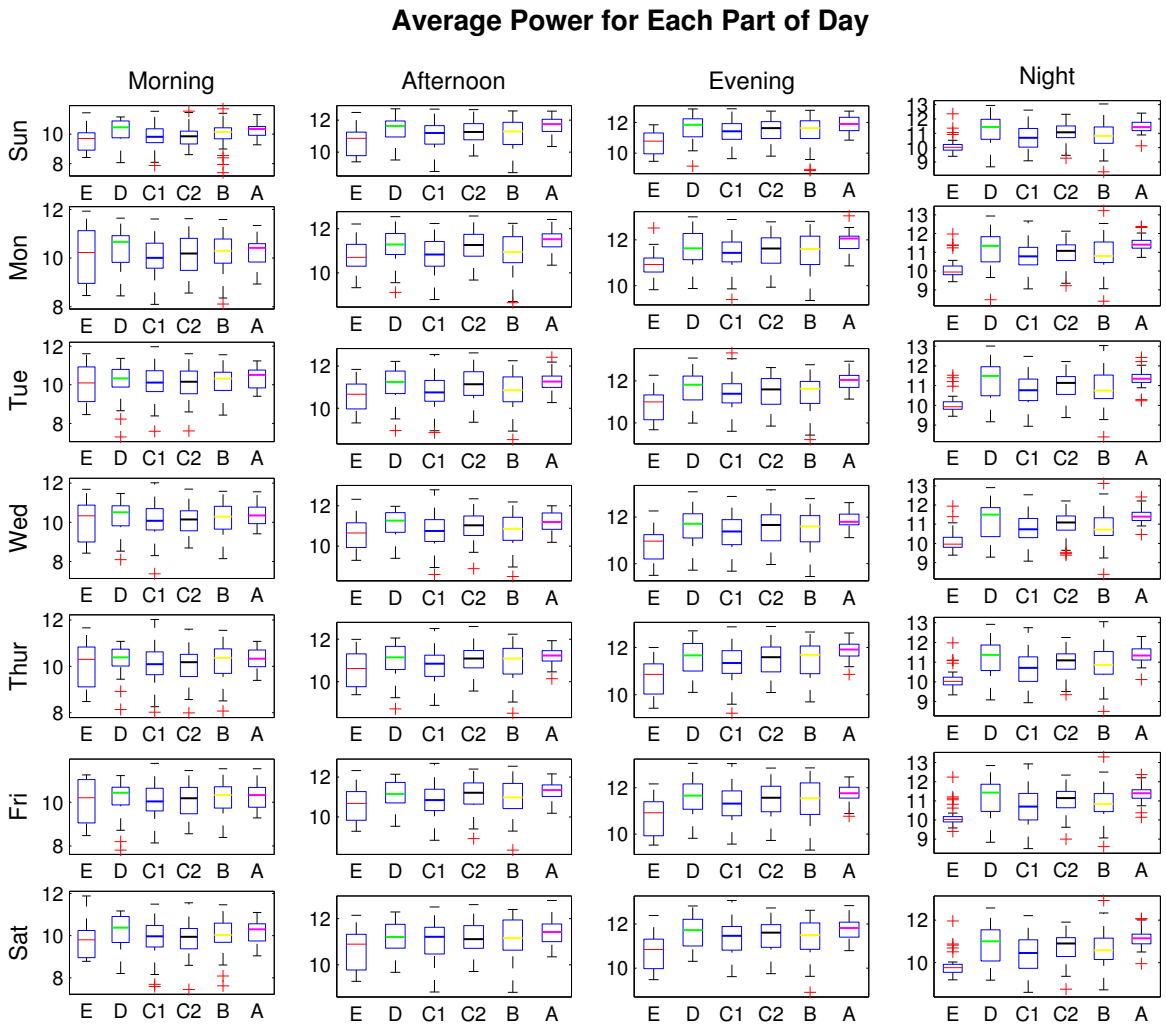


Figure 3.7

Figure 3.7, which compares socio-economic classes, shows again the same distribution as the previously computed features. Households of social grade E appear to use relatively little energy at night compared to households of other socio-economic groups, yet they seem to make up for it in the morning period, when their consumption is more akin to the other groups. Households of class A show the opposite pattern, using more energy than the others in the evenings, but normal amounts (compared to other classes) in the mornings.

Based on the observations thus far, it is likely that a classifier would have difficulty differentiating between socio-economic classes, particularly when discriminating between households of classes B or D and the rest. They have they have lower prior probabilities than C1 and C2, and do not look to have the extremes in energy usage that classes A and (particularly) E do.

Mean Weekday vs. Saturday and Sunday

In addition to looking at consumption features, ratios can also give insight into when a household is using its energy. Taking the ratio of the energy consumed

on a weekend day to weekdays, one can determine if a household is using proportionately more of its energy during the week or at the weekends. Households of social grades E, D and C2, whose chief income earners, when employed, are often either unskilled or manual labourers, or work in lower level positions within such industries as retail, hospitality and food services, are more likely to be required to work at the weekends than households of classes C1, B or A who, given their supervisory or managerial roles, are less likely to work at the weekends [34]. The feature is computed by taking the average energy consumed by a household on a Saturday and a Sunday, and dividing that by the average energy used during the week[2]. The features can be expressed as:

$$\text{Sat Ratio} = \frac{\text{mean total energy used on Sat}}{\text{mean total energy used from Mon to Fri}} \equiv \frac{\sum_{i=1}^{4} Sat_i}{\sum_{i=1}^{4} Mon_i + Tue_i + \cdots + Fri_i}$$

$$\text{Sun Ratio} = \frac{\text{mean total energy used on Sun}}{\text{mean total energy used from Mon to Fri}} \equiv \frac{\sum_{i=1}^{4} Sat_i}{\sum_{i=1}^{4} Mon_i + Tue_i + \cdots + Fri_i}$$

where Day_i is the total energy used on the i^{th} occurence of that day over the four-week period

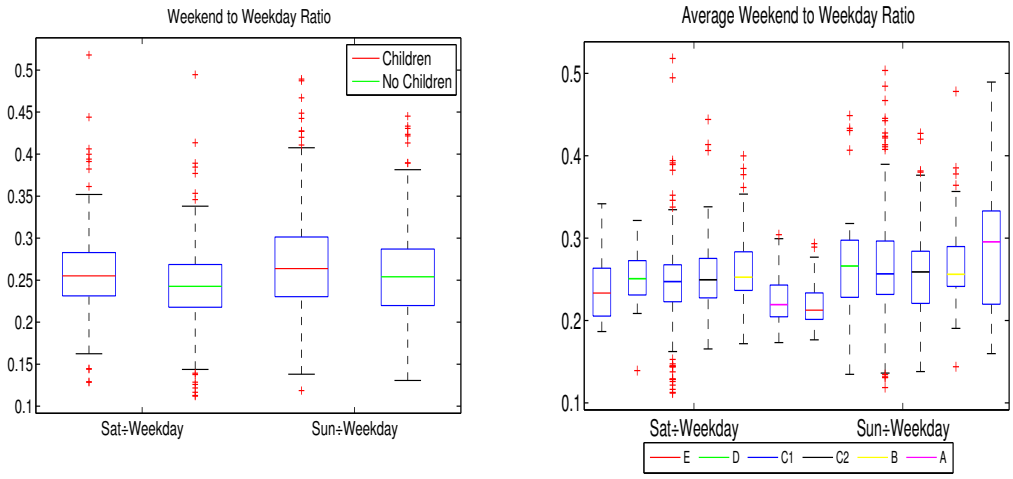


Figure 3.8: The ratio between the energy used on weekend days and the total used during the week, averaged over a four week period. This shows how the amount of energy used during the week compares to that used at the weekends.

After computing the ratio between weekend and weekday electricity consumption, classes seem to use similar proportions of their energy. And while Figure 3.8 suggests that households generally use more of their energy on Sundays than on Saturdays, this is independent of whether or not the household has children. Households of socio-economic class A are found to use a lower proportion of their total energy on Saturdays than other classes, but might use more on Sundays, whereas the opposite is true for households of class E. Again, there is the issue that households of classes D,C1,C2 and B are indistinguishable from one another based on the plot.

[2]Simply taking the total energy used on all Saturdays and Sundays and dividing it by the total used on weekdays would give the same result. However, the decision was taken to use the average because these values were conveniently available from previously computed features.

Variance on Weekdays

Thus far, the features that have been computed have been dependent on *how much* energy has been consumed. It is also worth considering what can be inferred from the volatility of a household's energy consumption over the course of a day. The assumption behind such a feature is that certain types of households will have more 'spikes' in their daily consumption. For example, households with greater disposable income (which is correlated with their socio-economic status) are more likely to own appliances that have high energy demands such as tubule driers and dishwashers.

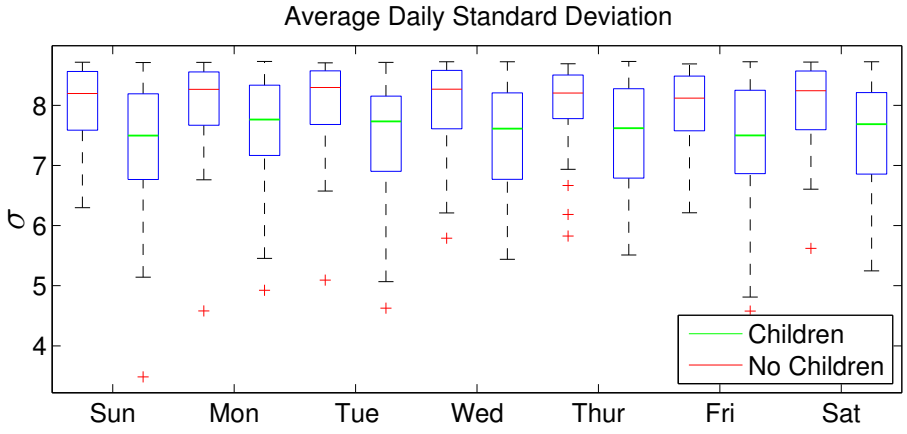


Figure 3.9: The variance of a household's daily consumption grouped by the day of the week and whether the household has children (C) or not (NC)

Although the average daily variance of households is volatile in and of itself, the results shown in Figure 3.9 indicate that the electricity use of households with children does tend to fluctuate more than those without children and therefore could be used to discriminate between households with and without children.

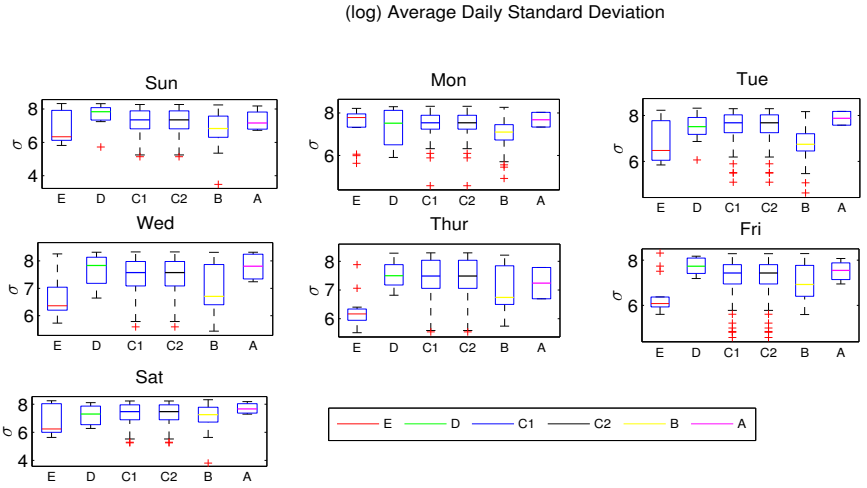


Figure 3.10: The variance of a household's daily consumption grouped by the day of the week and the Ipsos mori social grade

According to Figure 3.10, it is conceivable that the variance of a household's electricity consumption could be used to determine the socio-economic class of the household. For example, households of class E can be distinguished from the remaining classes based on the variance of their consumption on Thursdays and Fridays. On Tuesdays, households of class B stand apart from A, C2, C1 and D.

Correlation Between Weekdays

The final feature calculated from the consumption figures is the average correlation coefficient between one weekday and every other weekday. Each weekday was taken in turn, and the correlation between it and each subsequent weekday was calculated. This is done for each week separately and the averaged correlation of each pair of days is considered as one feature. Rather than using the 10-minute intervals, which appeared to be too granular to capture any covariance between days, electricity readings were summed into one-hour intervals. This is done because, while people might have certain tasks that they repeat daily (such as having a cup of tea), it is unreasonable to assume that it is always done during the same 10 minute period each day, but likely to be within the same hour.

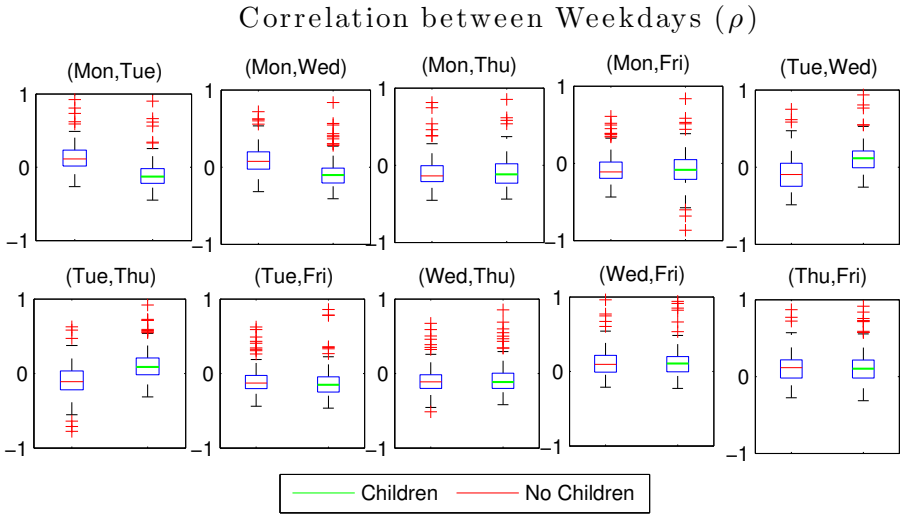


Figure 3.11: Average correlation coefficient between weekdays grouped by whether a household has children (C) or not (NC)

Looking at Figure 3.11, it appears that although the correlation coefficients are generally close to 0 (meaning there is no correlation), there are nevertheless differences between the two classes. Depending on which two days are being considered, the correlations of one class tend to be greater or smaller than that those of the others. For example, it would appear that households with children demonstrate a slightly higher correlation between their Monday and Tuesday electricity consumption patterns than those without. Whereas for socio-economic classification, as depicted in Figure 3.12, the correlation between days does not result in features that separate classes.

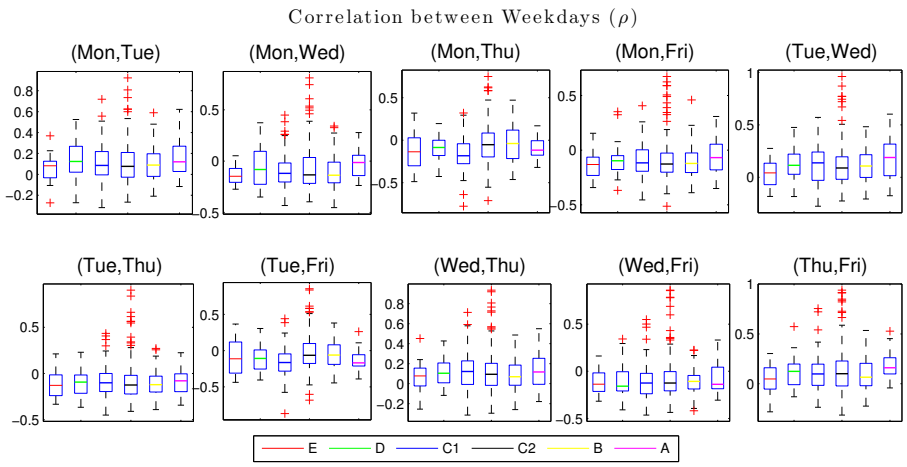


Figure 3.12: Average correlation coefficient between weekdays grouped by Ipsos social grade

3.4 Periodicity

Another approach used for feature extraction is to exploit the periodic consumption patterns exhibited by many households in order to search for temporal structures present in some classes but not in others. This method of feature extraction has been used successfully in previous studies involving forecasting and clustering. Methods outlined by Fabian Moerchen [28] for time series feature extraction are used to project each household's consumption into the frequency domain from which the most important frequencies are found. McLoughlin et al. [24] showed in their research that temporal structure is present in household electricity consumption data and can be used to characterise domestic energy demand.

Signal Smoothing

To extract the periodic structure of the data, each time-series can projected into frequency space by taking the Fourier transform (described in Section 3.3). However doing this, the Gaussian averaging operator was applied to each set of readings to filter noise whilst retaining the temporal structure of the data. Gaussian filtering (or Gaussian smoothing) is accomplished by convolving a time series with the Gaussian function. It can improve performance compared with direct averaging, as more structure is retained whilst noise is removed [35]. This is done because the time-frequency transformation used (the discrete Fourier transform method) has difficulty characterising small intervals of large electricity demand [36].

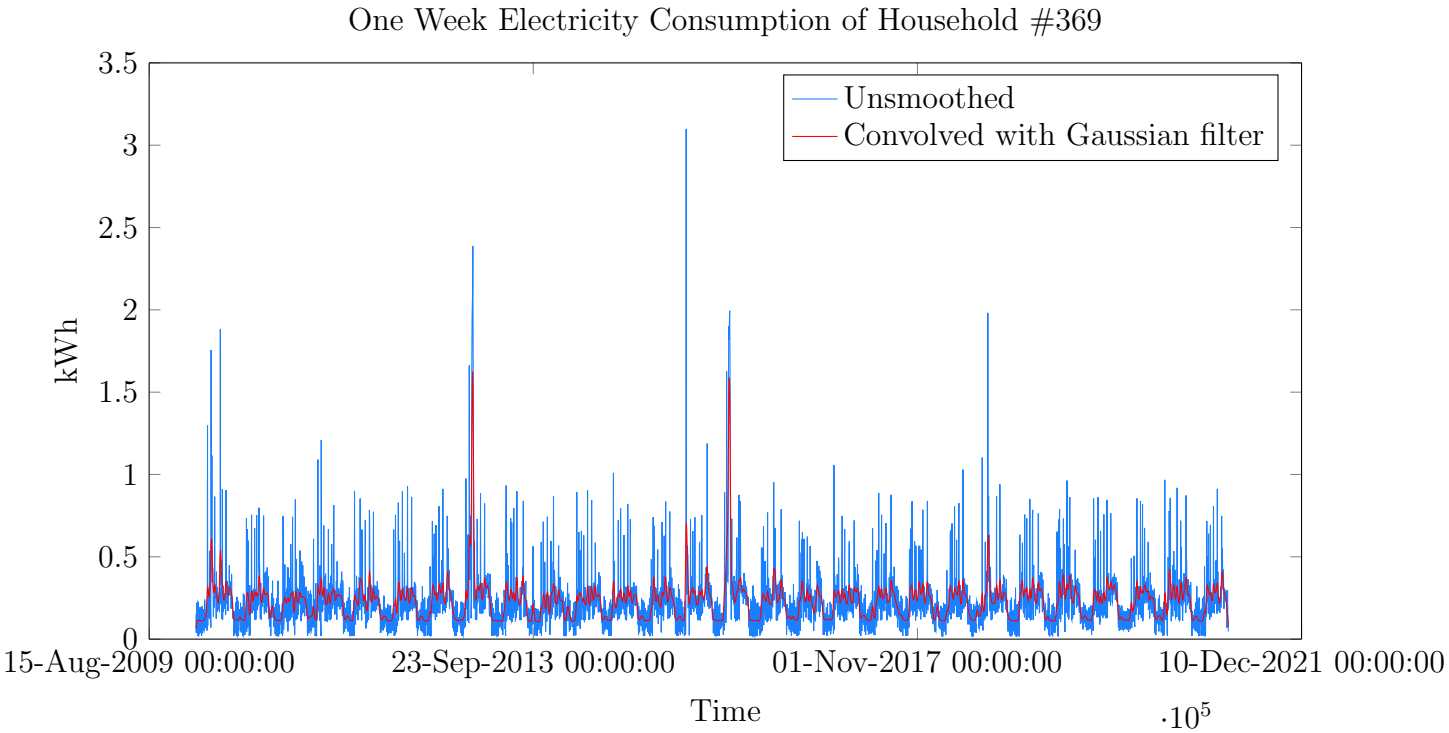


Figure 3.13: The electricity use of Household # 369 shows that households may have both a daily and weekly pattern. The clusters of peaks represent individual days while the regions without peaks are the indicative of night time. Additionaly, the large spikes are observed roughly every seven days, on either Saturdays, Sundays or both. After applying the Gaussian filter, the time series maintains its temporal structure however the sharp peaks are smoothed, which would not be handled well by the Fourier transform

Fourier Transform

For uniform samples $[f(1)..., f(n)]$ of a real signal $f(x)$, the *Discrete Fourier Transform* (DFT) is the projection of a signal from the time domain into the frequency domain by

$$c_f = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} f(t) \exp \frac{-2\pi i f t}{n}$$

where $f = 1, ...n$ and $i = \sqrt{-1}$. The c_f are complex numbers and represent the amplitudes and shifts of a decomposition of the signal into sinusoid functions [28].

Issues do present themselves when using this method. The Fourier transform measures global frequencies and the signal is assumed to be periodic. This assumption can cause poor approximations at the borders of the time series [28].

Energy Preservation

For l time series of length m, the DFT produces an $l \times m$ matrix C of coefficients, such that element $c_{i,j}$ is the j^{th} coefficient of time series i. In our case, since the number of households, $l = 519$, is small compared to the length of each time series, $m = 4032$, the number of coefficients must be reduced in order to minimise

redundancy, noise and computational time. According to Moerchen [28], the best subset of k columns is found by selecting those that optimize energy preservation E, defined as

$$E(f(t)) = \sum_{j=1}^{m} a_j c_j^2$$

where c_j is the j^{th} column and a_j is an appropriate scaling coefficient corresponding to signal $f(t)$.

Let I be a function measuring the importance of coefficient j on all values of l, and let $J_k(I, C)$ be a function that chooses a subset of $M = 1, ..., m$ of the k largest values of I. Moerchen [28] proves that $J_k(mean(c_j^2), C)$ is optimal in energy preservation.

The MATLAB fast Fourier transform function (fft) was used to find the discrete Fourier transform; the five best features were chosen based on the energy preservation method.

3.5 Dimensionality Reduction

Even though the success of a classifier is dependent on several variables, which may differ from one classifier to another, all classifiers are dependent on the quality of their input data. To achieve accurate results with the least amount of computational time, it is necessary to ensure that the minimal amounts of noise and redundancy is present in the input. This may involve dimensionality reduction, the process of identifying and filtering out as much irrelevant and redundant information as possible [37].

Different classification algorithms will be affected by overparameterisation in different ways. In the K-nearest neighbour classifier, additional features can largely affect the distance between two points. While redundant features (i.e., those that don't change the distance between points) would only influence computational cost, added noise to the system can impact the distance between points, likely in a negative way.

Like K-nearest neighbour, the need for feature reduction in logistic regression has less to do with removing redundancy than with reducing noise and computational cost. Logistic regression accounts for highly correlated features by lowering their weights. Uninformative features, however, would cause weights to be learned that do not improve the performance of the classifier.

Random forests are not as susceptible to the problem of overparameterisation as other methods. When training each tree, since the 'best' features will be branched on towards the top of the tree, pruning could be used to limit the size of each tree (thus avoiding overfitting). An issue would only start to arise when the number of redundant or noisy features become much larger than the number of good features. This is because, when training a tree, a random subset of features is selected as each branch is created. If the number of 'bad' features is much larger than the number of 'good' ones, then the probability of choosing a subset where no good features are present becomes significant.

Dimensionality reduction can usually be characterised as one of two tasks: *feature selection* and *feature transformation*. Feature transformation methods

involve performing a transformation of the data (such as a rotation or projection) to create a new set of features (of smaller size) that has more descriptive power than the original set. A commonly used example of this is *principal component analysis* (PCA) which finds a set of orthogonal unit vectors that point in the directions of greatest variance of the data. The features are given by projecting the data onto this basis. While these sorts of methods are popular and do tend to perform well, the resulting features are usually not interpretable [38].

It might be of interest to see which features are most responsible for differences between classes. Therefore, instead of using feature transformation methods, feature selection is used to find a subset of features for which a classifier achieves its best performance. There exist numerous methods for performing feature selection, such as nested subset methods, filters or direct objective optimisation [38], as well as adaptive boosting [39].

Sequential floating selection (SFS) [40] was used to find the optimal set of features. SFS is a greedy algorithm that works by: starting with an empty list and a set of n features, sequentially considering each feature and assessing its impact on a given evaluation score (e.g. constructing a classifier that takes only one feature and finding its error). It chooses the feature that scores best and adds it to the list. It then goes through each of the remaining $n-1$ features that have not been added to the list, and assess their performance in combination with the features already added to the list (e.g. a constructing a classifier that takes two features and evaluating it). It find the feature that improves performance most and adds it to the list. This is repeated until the list is full [41]. A superior method, *sequential forward floating selection*, has been proven to perform better [40], which backtracks after a new feature is included to solve the *nesting* problem, it proved inefficient to implement for the multi class and was therefore not used.

3.5.1 Implementation

Since it is not necessarily the case that the best features are the same for each classification problem, or even for each classification algorithm, the best features are found for each classifier irrespective of the others. The figure of merit for each, which is optimises the classifier is found by using cross-validation and training a classification model with training data and then evaluating it on a validation set. If at any stage, the feature being considered improves the figure of merit, then the feature will be added to the set of 'kept' features.

Different evaluation scores (cost functions) are used depending on the classifier. In the k-nearest neighbours classifiers, the *mincost* is used, which is the predicted label with the smallest expected misclassification cost. The expectation is taken over the posterior probability, and cost as given by the Cost property of the classifier (a matrix). The loss is then the true misclassification cost averaged over the observations. For the random forest implementation, the cumulative misclassification probability of the entire ensemble is used as the cost to evaluate combination of features. In the case of logistic regression, the deviance of the fit is used. These methods were used for two reasons, firstly because efficient implementations exist with MATLAB's stats toolkit, and secondly, they produced the sets of features that performed best on when tested on a validation set.

Chapter 4

Models

4.1 Overview

There are several classification algorithms that can be used to perform supervised learning tasks and vary in their computational complexity, implementation and assumptions that they make about the distributions of the data [10]. Three well known methods are used to classify the data: Logistic regression, random forrest and k-nearest neighbour.

All three methods are examples of discriminative classifiers. The discriminative approach is appealing in that it it directly models $p(y|\mathbf{x})$. Also, density estimation for the class-conditional distributions is a hard problem, particularly when $\mathbf{x}$ is high dimensional, so if we are just interested in classification, then the generative approach may mean that we are trying to solve a harder problem than we need to[42].

4.2 Logistic Regression

For a binary classification problem $y \in \{0, 1\}$, such as discriminating between households with children ($y = 1$) and households without ($y = 0$), the logistic regression model learns a weight vector $\mathbf{w}$ such that given some new household with feature vector $\mathbf{x}$, the posterior probability of that household being in class, $p(y = 1|\mathbf{x}) = g(\mathbf{x}; \mathbf{w})$ where $g(x)$ is the logistic (or sigmoid) function.

$$g(\mathbf{x}; \mathbf{w}) = \sigma(\mathbf{x}; \mathbf{w}) = \frac{1}{1 - e^{-(b+\mathbf{w}\cdot\mathbf{x})}}$$

There are numerous advantages to using logistic regression for the household classification task. Firstly, logistic regression is interpretable. After the model has been trained and the weight vectors established, they can be used to determine how important each feature is to the classifier. Secondly, the confidence of a prediction can be inferred, resulting in interpretable results. There are, however, also drawbacks to logistic regression. Since the maximum likelihood function does not have a closed form solution, an iterative process must be used instead to learn the weights, which is not guaranteed to converge.

4.2.1 Multi-class Logistic Regression

To extend the problem of logistic regression to the multi-class case, often times the the *softmax* is used as a generalisation of the logistic function (σ), the predicted class of an instance is then given by

$$P(y = Y_i|\mathbf{x}) = \frac{\exp^{-(b_i+\mathbf{w}_i\cdot\mathbf{x})}}{\sum_{j=0}^{J}\exp^{-(b_j+\mathbf{w}_j\cdot\mathbf{x})}}$$

Although this is a valid method of classifying the data, it fails to acknowledge the ordinal property of the classes and assumes the data to be nominal. Ideally we would be able to build a model that exploits the fact that some classes are more similar than others. For example, if the true label of a household is B, then we would rather missclassify the instance as A or C1 than as D or E. Luckily, ordinal logistic regression (or ordered ligit) can be used to build a model that incorporated the ordering of the classes.

McCullagh's proportional odds model [43] is a variation of a generalised linear model (glm) where the dependant variable is thought of as being continuous, but is recorded ordinally. It can be thought of as asking asking a linear model to tell you what range a dependent variable is in (as opposed to an exact value). The regression model is then given by:

$$\begin{aligned}
logit(p_1) &= \log(\frac{p_1}{1-p_1}) = b_1 + \mathbf{w}\cdot\mathbf{x} \\
logit(p_2) &= \log(\frac{p_1+p_2}{1-p_1-p_2}) = b_2 + \mathbf{w}\cdot\mathbf{x} \\
&\vdots \\
logit(p_6) &= \log(\frac{p_1+p_2+p_3+p_4+p_5+p_6}{1-p_1-p_2-p_3-p_4-p_5-p_6}) = b_6 + \mathbf{w}\cdot\mathbf{x}
\end{aligned}$$

This model relies on the *proportional odds assumption*, which is that the $\mathbf{w}$s are independent of the classes (hence they have no subscripts). This translates to the assumption that the weights are the same for each cutoff, but rather the classes have different intercepts b, in contrast to multinomial logistic regression (where the dependent variables are assumed to be nominal) which learns a new set of weight parameters for each cutoff point. A further description of regression models for ordinal data in the context of machine learning is given by Herbrich et. al. [44].

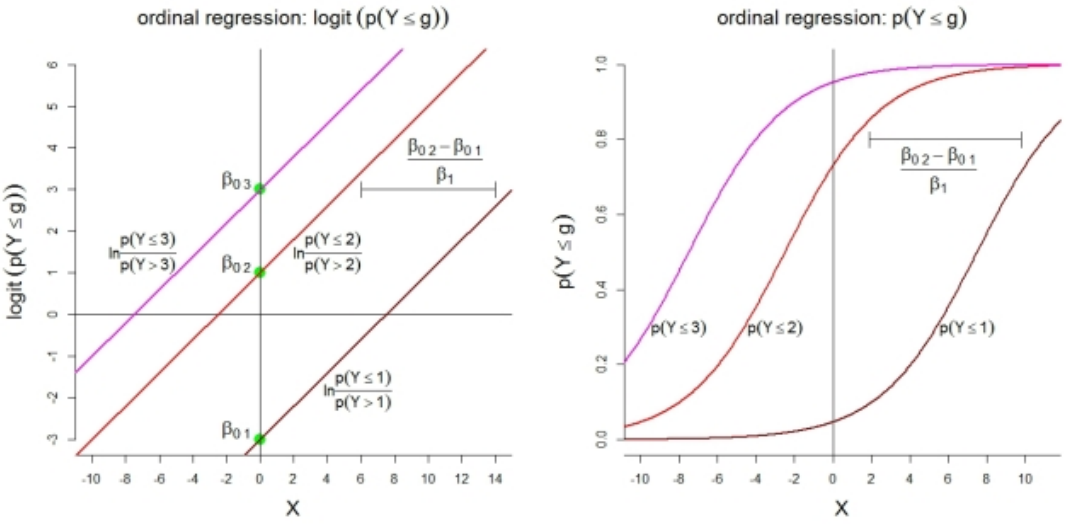


Figure 4.1: The proportional odds model assumes that the regression function for different response categories are parallel on the logit scale. Figure obtained from: http://www.datavis.ca/courses/grcat

4.2.2 Implementation

To build the binary logistic regression classifier, MATLAB's fitglm tool which fits a generalised linear model to the data. Logit link fuctions and a binomial distribution are passed as parameters to the glmfit function which produces a logistic regression model. For the case of multi-class classification, two models were created. One using McCullagh's proportional odds model (treating the data as ordinal) and one using multinomial logistic regression (treating the data as nominal). Both methods were implemented in MATLAB using the mnrfit tool.

4.3 Random Forest

Random Forest is a classification method that grows an ensemble of decision trees from a set of training instances and determines the class of a new instance by allowing the trees in the ensemble to vote on the most popular class. For N training sets and M features, each tree is grown by:

- Randomly sample n training instances from the N training with replacement (this will be the training sample to grow the tree).
- At each node, selecting m features at random (where $m < M$), the best of which is used to split the node.
- The trees are grown to the largest possible size (no pruning takes place).

A new instance is then classified by running it through each tree, allowing each of the trees to assign the instance a class. The predicted class of the test instance is then given by the vote of each tree.

Although (in contrast to building a single decision tree), it is not easy to visualise a random forest, it is still possible to gain an estimate of the variables

that are most important for classification and can be used on data sets with a large number of features (see Section 3.5). Random forests have been shown to perform particularly well on unseen data compared to other classification methods as they avoid overfitting by only ever looking at a random subset of features and data [45].

4.3.1 Implementation

MATLAB's built-in treeBagger class was used to build the random forest. Because bootstrap aggregation is used to randomly sample the training data, the out-of-bag (oob) estimates were used to optimise the model's parameters instead of using cross-validation. The parameters to optimise are the number of trees in the forest, and number of features m to consider for splitting each node. It was found that for both classification tasks, optimum number of trees in the forest is 13, and that evaluating $\frac{M}{2}$ randomly selected at each node gives the best oob error.

4.4 K-Nearest Neighbour

K-nearest neighbour is a fundamental method for classification as it is intuitive and requires little *a priori* knowledge about the data. It is a non-parametric model that classifies an unlabelled input by finding the K-nearest training points in feature space, using the classes of the nearest points to predict the class of the unlabelled point [46].

4.4.1 Implementation

MATLAB's fitknn tool was used to build a nearest neighbour classification model and the optimum parameters were found using 5-fold cross-validation. The parameters to find were the distance measure, search method and k (the number of neighbours). Using the 5-nearest neighbours and euclidean distance gave the best results in cross-validation.

Chapter 5

Results

This section discusses the quantitative evaluation methods used to determine the potential for each of the classifiers to reveal household characteristics, and then analyses the results from training and running classifiers.

5.1 Evaluation Methods

For each classifier, a *confusion matrix* (CM) is produced using the MATLAB tool `confusionmat`, which, for a K class classification problem, returns a $K \times K$ matrix, where each element (i, j) contains the number of times an instance of class i has been classified as j. The diagonal elements of CM contain the number of instances of households that have been classified correctly for each class. [47]

The accuracy of a classifier is defined as the sum of the diagonal elements of CM, divided by the total number of samples, S.

$$ACC = \frac{\sum_{i=1}^{K} CM_{i,i}}{S}$$

This is compared to the accuracy of performing a random guess (RG), which assigns a household to one of the K classes at random.

$$ACC_{RG} = \frac{1}{K}$$

As a benchmark comparison for the classifiers, the baseline accuracy is calculated by assigning all households to the most probable class (MPC), based on the class with highest prior probability.

$$ACC_{MPC} = \frac{argmax(S^K)}{S}$$

where S^K is the number of samples from the test data that are in class K.

For socio-economic classification problems, the ordinal structure of the classes should also be taken into account (i.e., it is worse for our classifier to predict a household of social grade B as D, than it is to predict B as C1 or A). Therefore, the *accuracy within n is also presented*, where n is the number of neighboring classes.[48].

Particularly for unbalanced classes, reporting the accuracy alone is not satisfactory in determining the quality of a classifier, an obvious and well known example being: construct a classification problem where 99% of instances are in class A and only 1% are in class B. A classifier that simply predicts all new data as class A would be correct 99% of the the time, but would still not be a good classifier.

A widely applied method for evaluating a classifier is to compute the *true positive rate* (TPR) and *true negative rate* (TNR). The TPR gives the proportion of positives that are correctly identified as being positive, while the TNR gives the proportion of negatives that are correctly identified as negative.

$$TPR = \frac{TP}{TP + FN} \qquad\qquad TNR = \frac{TN}{TN + FP}$$

From these statistics, it is common to plot an ROC curve. This is a plot of the TPR against the *false positive rate* (FPR), which is defined as 1-TNR. The evaluation criterion (the area under the ROC curve) is preferred over the accuracy, particularly when considering unbalanced classes since the impact of skewness can be analysed using the trade-off between TNR and FPR at different thresholds. [49].

This is not as straight forward for random forests and Knn as it is for probabilistic classifiers such as logistic regression. Probabilities can, however, be generated from the classifier results. For random forests the decision boundary may be the ratio of the number of trees that vote in favour of assigning an unseen instance to class 1 and the total number of trees. In Knn it is the number of nearest neighbours that are of class 1 divided by the total number of nearest neighbours.

In binary classification problems, such as evaluating whether households have children, the concept of 'positives' with relation to 'negatives' is quite straight forward. This is not the case, however, with multi-class problems, where the concepts are clouded. Beckel et al., binarised their groups and used a one-verses-all approach when constructing their ROC curves [23, 10]. A similar method was employed in this project, combining pairs of neighbouring classes and then labelling them as positive. The results were evaluated against the remaining classes.

The final metric worth discussing here is the Matthews correlation coefficient (MCC), a value between -1 and +1 that represents the correlation between the predicted and true outcomes of a binary classifier. An MCC of -1 indicates that that there is perfect anticorrelation between the predicted and true class, while a value of +1 suggests a perfect classifier, and a value of 0 means the model is no better or worse than a random guess. MCC was selected because it gives a value to the performance without inflating the imbalances among class sizes[1][50].

Let X, Y each be an $S \times N$ matrix where S is the number of households and N is the number of classes. $X_{s,n} = 1$ if a sample s is predicted to be the n^{th} class and 0 otherwise. $Y_{s,n} = 1$ if a sample s belongs to the n^{th} class and 0 otherwise. The covariance of X and Y can then be written as

[1] as opposed to, for example, the F1 score, which does not take into account the true negatives. In our case is the number of correctly identified households with children.

$$cov(X,Y) = \frac{1}{N}\sum_{s=1}^{S}\sum_{n=1}^{N}(X_{s,n} - \bar{X_n})(Y_{s,n} - \bar{Y_n})$$

where $\bar{X_n}$ and $\bar{Y_n}$ are the means of the n^{th} columns of X and Y, respectively. The MCC is then defined as,

$$MCC = \frac{cov(X,Y)}{\sqrt{cov(X,X) \cdot cov(Y,Y)}}$$

For binary classification this can be interpreted as,

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

While the MCC is not commonly used when assessing binary classifiers, based on the definition of the MCC above, it can be extended to multi-class problems, as described by Gorodkin[51]. If C is the confusion matrix of a classifier, then the MCC is given as,

$$MCC = \frac{\sum_{k,l,m=1}^{N} C_{k,k}C_{m,l} - C_{l,m}C_{k,m}}{\sqrt{\sum_{k=1}^{N}\left[\left(\sum_{l=1}^{N} C_{l,k}\right)\left(\sum_{f\neq k,g=1}^{N} C_{g,f}\right)\right]} \cdot \sqrt{\sum_{k=1}^{N}\left[\left(\sum_{l=1}^{N} C_{k,l}\right)\left(\sum_{f\neq k,g=1}^{N} C_{f,g}\right)\right]}}$$

5.2 Feature Selection

As explained in Section 3.5, SFS was used to determine the features that are of greatest value for each classifier. However, after running SFS multiple times, we noticed that the features found to be 'optimal' for each of the classifiers were not always the same (even after cross-validation). This was especially evident in random forest[2]. As such, the feature selection algorithm was run multiple times and the set of features (for each classifier) that appeared most often was used. To evaluate the feature selection method, two additional sets of features were constructed (one for each classification problem) by choosing features based on how they appeared, by visualisation alone, to separate the classes in Section 3.3. This was used as a baseline for comparison because it allowed us to exploit our domain knowledge, rather than have to rely on a random guess.

All classification models were evaluated using the features found by SFS as well as the features found manually (labeled as MAN). The lists of features used for each classifier are shown in Appendix A.

5.3 Classification Results

This section illustrates the results obtained by testing each model on unseen data, as outlined in Section 5.1. First, the results of discriminating between

[2]because random forests rely on random feature subset selection.

households with and without children are presented, followed by the results of the socio-economic classifiers. These are discussed to determine which classifier and set of features performs best on the classification tasks.

The same training and test sets were used for each classifier to ensure that the results were fair. While the training set was the same as that used in cross-validation to optimise the classifiers, the test data was entirely unused up to this point.

5.3.1 Children vs No Children

Child vs. No Child Confusion Matricies

		SFS		MAN	
		Predicted Class			
Log Reg	Actual Classal	59	11	56	12
		6	28	15	21
Random Forest		57	13	62	8
		7	27	14	20
Knn		62	8	61	9
		10	24	16	18

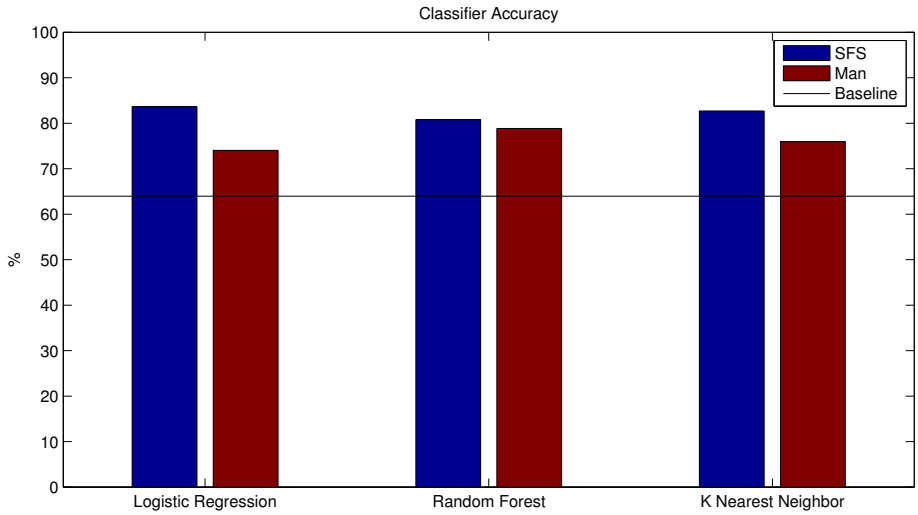


Figure 5.1: The accuracy of each classifier when used to classify unseen data defined as the number households classified correctly as a percentage of the total households in the test set. SFS classifiers are those built using features determined to be best by the sequential forward feature selection method while, MAN are the features found by visually determining which features appear to best discriminate between the data

The results in Figure 5.1 show the accuracy of each classifier built to discriminate between households with and without children. All the classifiers performed better than the baseline accuracy (in this case, the accuracy obtained by classifying all households as being without children). The logistic regression model, built using features selected by SFS, predicted the greatest percentage of households correctly.

Figure 5.2 illustrates the Matthews correlation coefficients (MCC) of each of the classifiers. As discussed in Section 5.1, the MCC is a more suitable performance measure than accuracy because it 'rewards' correctly identifying samples from the under-represented class and 'punishes' predictors that are strongly biased [10]. Indeed, all the classifiers took on positive values, meaning that the classifiers were less inclined to misclassify a household regardless of which class had a higher prior probability.

While the logistic regression classifier using the SFS features gave both the highest accuracy and MCC, the performance was not significantly better than the random forest and Knn methods (also using SFS). What was distinctly noticeable was that all classifiers performed better when trained with features found by SFS as compared with manually selected features (MAN). The accuracy of all three classifiers that used SFS came in above 80%, whereas those trained on manually selected features were below 80%. The distinction was even more pronounced in each of the classifier's MCC's. All classifiers using SFS features had an MCC of at least 0.59, while the best of the manually selected features obtained an MCC of 0.5.

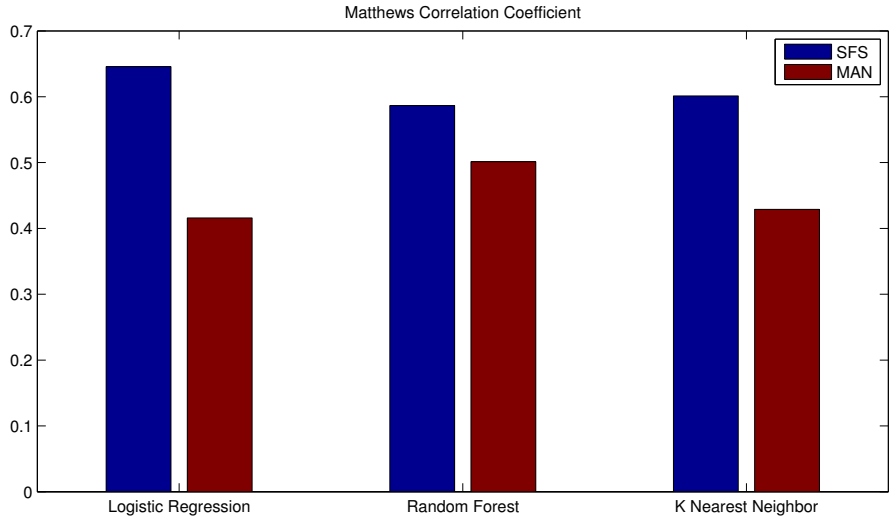


Figure 5.2: The Matthews correlation coefficient of each classifier when used to classify unseen data. SFS classifiers are those built using features determined to be best by the sequential forward feature selection method, while MAN are the features found by visually determining which features appear to best discriminate between the data

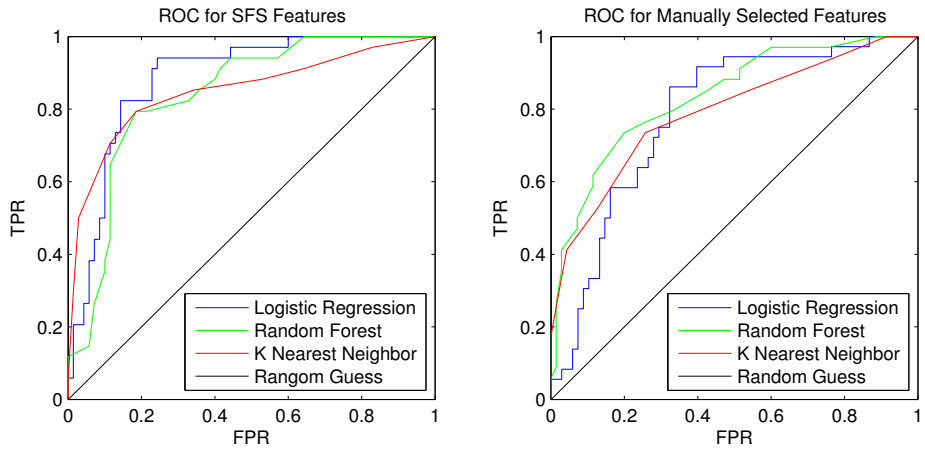


Figure 5.3: Receiver operating curves that show the trade-off between true positive rate and false positive rate.

The ROC curves in Figure 5.3 show how the TPR changes when varying the FPR. Again, it can be seen that SFS features generated better results than MAN features. All three classifiers trained on SFS features could identify 70% of households without children, whilst only mislabeling 11% of those with children, whereas the classifiers trained on MAN features would have needed an FPR of at least 17% to have achieved the same TPR. Again, the logistic regression classifier gave the best performance (based on the area under the curve). Looking at the Knn classifier, the TPR only had a steep gradient when the FPR was very low. But as FPR limitations on the number of false positives that are allowed was relaxed, logistic regression and random forest started to outperform.

Children vs No Children

	Accuracy (%)		**MCC**		**ROC Area**		**TPR**		**TNR**	
Classifier	SFS	MAN	SFS	MAN	SFS	MAN	SFS	MAN	SFS	MAN
Log Reg	83.6538	74.0385	0.6457	0.4159	0.8840	0.7831	0.8429	0.8235	0.8235	0.5833
Random Forest	80.7692	78.8462	0.5866	0.5012	0.8397	0.8357	0.8143	0.8857	0.7941	0.5882
KNN	82.6923	75.9615	0.6013	0.4289	0.8502	0.7905	0.8857	0.8714	0.7500	0.5294

Finally, the table above shows the TPR and TNR for each of the classifiers along with the other previously discussed statistics. The logistic regression classifier achieved the greatest accuracy, MCC and area under the ROC, although it did not have the highest TPR. From this one can interpret that the logistic regression classifier was not overfit to the training data[3]. In contrast, the Knn classifier was able to correctly identify 88.6% of the households without children.

Looking at the TNR of each of the classifiers, Knn performed worst when using both SFS and MAN features. Since it had the highest TPR, it is obvious that the classifier relied more on the prior probability than the others. This is intuitively obvious: if there are two clusters that overlap in feature space (i.e., are not separable) and one cluster has more points, then any new point in the same space will see more neighbours of the class that occurs more frequently.

From looking at the TNR, it can also be seen that the models built using MAN features relied more on the imbalances in the sample sizes, whereas the

[3]because the positive class also had a larger sample size.

models that trained using SFS features exploited differences in the values of the household attributes to make more informed decisions about a household class.

This outcome indicates that information can be inferred from smart meter data.

5.3.2 Socio-Economic Group

While the results from the previous section showed that it is indeed possible to discriminate between households with and without children, the classifiers built to determine a household's socio-economic status were not as promising. The confusion matrix generated by each classifier, as shown in Table 5.4, gives some insight into how the individual classifiers predicted unseen data. The first thing to notice is that the ordinal logistic regression models were only slightly better at predicting a household's socio-economic status than biased random guesses. They predicted almost all test instances as either class C1 or C2, the two classes with the highest prior probability based on the sample population ($p(C1) = 0.38$, $p(C2) = 0.25$).

The nominal (multinomial) logistic regression classifiers were not much more accurate, although they were not as heavily biased as the ordinal regression model towards the most probable classes. The weights learned by the multinomial model can confirm that the proportional odds assumption was not satisfied. Appendix A shows the table with the weights learned. If the proportional odds assumption had been upheld, then the weights across each row would have been similar (other than the bias w_0).

Socio-Economic Class Confusion Matricies

		SFS						MAN					
		Predicted Class											
Ordinal Log Reg	Actual Class	1	0	4	1	0	0	0	0	2	4	0	0
		0	0	3	5	0	0	0	0	1	7	0	0
		1	0	7	18	3	0	0	0	3	26	0	0
		0	0	4	32	4	0	0	0	2	38	0	0
		0	0	3	11	1	0	0	0	1	14	0	0
		0	0	0	6	0	0	0	0	0	6	0	0
Nominal Log Reg		4	0	0	2	0	0	2	0	0	4	0	0
		1	0	1	2	3	1	1	0	0	3	4	0
		0	0	11	16	1	1	1	0	9	16	2	1
		2	1	4	29	4	0	3	1	4	31	1	0
		2	0	1	9	2	1	0	0	0	14	1	0
		0	0	1	3	1	1	0	0	1	4	0	1
Random Forest		4	0	0	2	0	0	4	0	1	1	0	0
		0	0	1	2	5	0	0	1	1	4	2	0
		0	1	18	9	1	0	0	1	20	7	2	0
		1	1	5	26	5	2	2	1	6	25	4	4
		0	0	1	7	6	1	1	0	3	6	4	1
		0	0	1	3	2	0	0	0	3	1	1	1
Knn		4	0	0	1	1	0	3	0	2	1	0	0
		0	1	1	4	2	0	1	1	0	2	4	0
		0	1	16	9	1	2	1	0	20	6	2	0
		2	1	8	24	5	0	1	2	6	28	2	1
		1	0	2	9	3	0	1	0	0	7	7	0
		0	0	1	3	2	0	0	0	3	0	2	1

Figure 5.4: Each confusion matrix shows the number of households that were classified as a given class, compared to their true class. Each row represents the true label of an instance while the columns represent the predicted label, i.e element (i, j) represents the number of households that were classified as j but who's true class was i. The rows and columns go in order of increasing social grade, so the 1^{st} row/column represents class E, the second represents class D ect.

From the confusion matrices it can also be seen that all of the classifiers gave a very low probability of a household being in class D. This is not surprising as the figures outlined in Chapter 3 suggested that it should have been difficult to distinguish between households of socio-economic classes D, C1 and C2. Since D had a much lower prior probability than the other two, a probabilistic classifier would likely have assigned an unseen instance to either class C1 or C2.

Looking at Figure 5.5, it can be seen that, although the accuracy of the classifiers starts to quickly increase as more neighbouring classes are considered positive, the benchmark accuracy also increases rapidly. This is because the most probable classes are in the middle of the ordering and are therefore more likely to be within the n nearest neighbours.

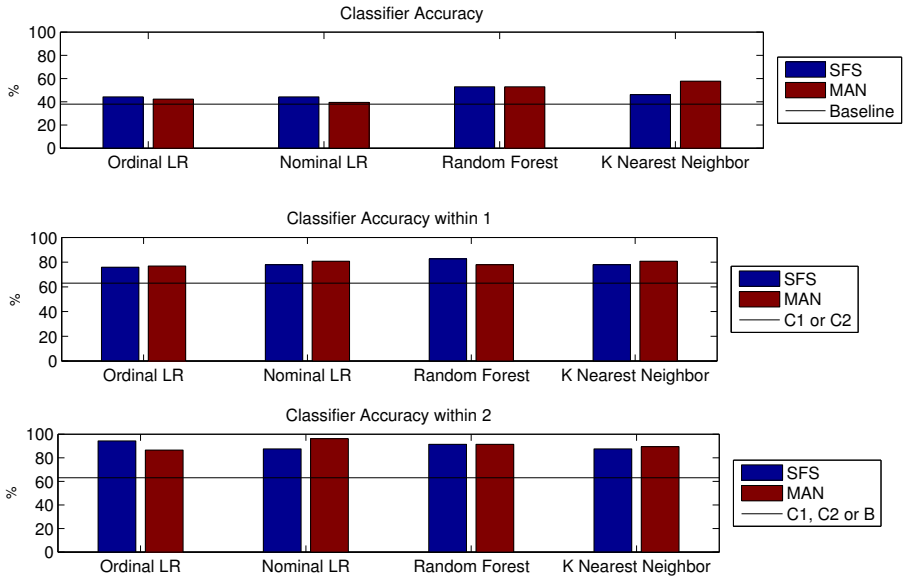


Figure 5.5: The accuracy of each classifier is given by the sum of the diagonal elements of its confusion matrix. The accuracy with n is the sum of the diagonal elements in addition to the elements up to n columns to the left and right of the diagonal.

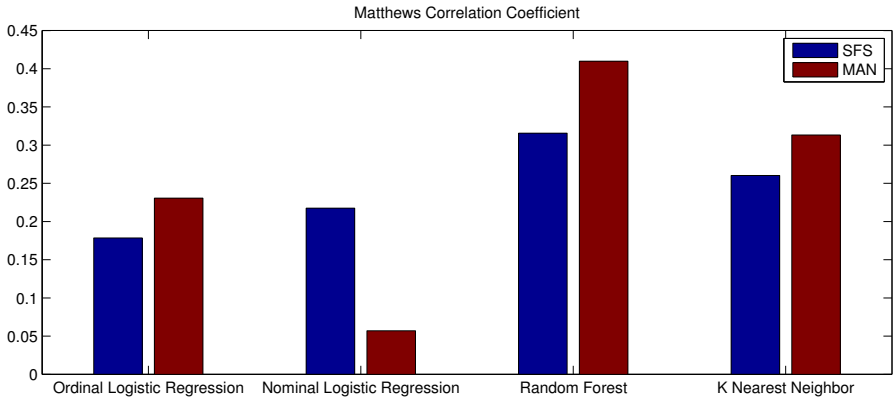


Figure 5.6: The Matthews correlation coefficient of each classifier when used to classify unseen data. SFS classifiers are those build using features determined to be best by the sequential forward feature selection method while MAN are the features found by visually determining which features appear to best discriminate between the data

Figure 5.6 shows the MCC's of each classifier. MAN features tend to do better than SFS here. SFS is a greedy algorithm and, particularly for multi-class problems with limited amounts of data to perform validation on, a backwards step such as that used in SFFS could render a more optimal set of features. Random forests are less affected by this as they determine which feature is best at each stage of being grown. The MCC treats the data as nominal and penalizes all misclassifications equally. A modification to the MCC that accounts for the ordering of the data would give a better evaluation of each of the models, however no such method was found in the literature.

The final evaluation metrics computed were ROC curves, which we generated by grouping classes A and B, C1 and C2, and D and E together. Each of these groups was then evaluated against the other two. As shown in Appendix A, the classifiers' performances were only useful when trying to extract C1 and C2 households and were otherwise perverse.

While this does not prove that socio-economic information can't be extracted from smart meter data, it does show that the problem is non-trivial.

Chapter 6

Discussion

In this chapter the important results of the models are discussed and summarized and we try to draw conclusions about the household classification problem.

There are several major discussion points that could be outlined on the basis of the evaluation of the experimental design and so this chapter is split into 3 sections:

- Features and feature selection
- Classifiers
- Comparison to related work

6.1 Features

A large portion of the time spent on this project was dedicated to gaining an understanding of the HES dataset and constructing a set of features that captured as much of the variance of the data in as low of a dimensionality as possible, whilst still being interpretable. Given that one of the two classification problems was accomplished with high accuracy and relatively low error, we can conclude that it is possible to build such a set of features. Below are some of the key observations and questions regarding feature extraction and selection that were encountered in this project.

1. *Time-frequency transformation does not capture differences in the periodic structure between household types*

   Despite numerous attempts to extract informative frequency amplitudes from time-series, we were unable to isolate features that could reliably distinguish between classes of households. Among the various time intervals considered were 4-week periods, weeks and individual days. Time-series where the weekend period had been removed were also examined to see if there were any patterns in weekdays that were being disrupted by weekends. Although taking the fft did identify periodic structures in the households' energy consumption, these could not be attributed to any one class relative to the others.

2. *Sequential forward selection provides a better method of classification than relying solely on domain knowledge*

   As shown in the results, in the problem of discriminating between households with and without children, the classifiers that relied on features selected by SFS performed better than those using features selected by hand on almost all accounts. The performance metric with the largest gap was the true negative rates. Classifiers using SFS features had, on average, 28% better results.

   In the socio-economic classification problem, though, SFS performed considerably worse than choosing features based on human intuition. Interestingly, however, the classification accuracy appeared relatively comparable. The difference was seen in the MCC's, which were markedly lower for most of the classifiers built on SFS features, negating any merits found elsewhere. One possible explanation for the discrepancy was that the sample sizes of the classes were small, so more data would allow the models to determine the best features with more confidence. But the dataset used by Beckel et al. and McLaughlin et al. was 18 times larger than than the HES dataset (and recorded houses for longer periods of time), and even so, they were no more successful in determining socio-economic class from their sample data.

   Another explanation might be been that, because SFS is a greedy algorithm, it is subject to the 'nesting effect'. In other words, once a feature had been added it couldn't be removed [38]. This would need to be further investigated in order to confirm or deny.

   The features selected by SFS were dependent on the cost function applied to penalise the classifiers. Numerous cost functions were assessed for each of the classifiers, including the misclassification rate, mean absolute error and the false positive rate. When testing the veracity of the features obtained from each cost function on the validation set, it was found that using a cost function that indicates some form of confidence or predicted future error gave better performance than ones that relied on the performance on a validation set.

3. *The optimum set of features is different for each classifier, even when tackling the same classification problem*

   During the parameter optimisation stages, SFS was implemented such that the training set was randomly split into five validation sets each time the algorithm was run. This resulted in, especially for inferring socio-economic group, the SFS method learning a new set of 'best' parameters each time the SFS algorithm was run. While this could be attributed to the limited sample size and a larger set of data would have given more stability in the results, it is more likely a result of an oversight in the implementation. The cross-validation method used in running SFS didn't check that all classes were represented proportionately in each validation set. Running the algorithm numerous times for each classifier and choosing the sets that occurred most frequently mitigated the effects, however this will have affected the results.

6.2 Classifiers

While the majority of the time was spent understanding the data and extracting features, the classifiers played a crucial role in the outcome of the project. The three classifiers, logistic regression, random forest and Knn are discussed here.

6.2.1 Logistic Regression

1. *Outperforms random forest and Knn in discriminating between households with and without children*

   With the greatest area under the ROC curve, along with highest accuracy and MCC, the logistic regression classifier is concluded to be the best-performing model for this specific binary classification task. While Knn was able to better identify the childless households, it fell short in identifying the households with children. The logistic regression model did a better job at identifying households that had children present.

   It is possible that the underlying reason for the positive results was that households with children will tended to have more occupants, which is correlated to the amount of energy consumed by the household[19]. However, after testing the correlation between the misclassified samples and the number of occupants of a household, no correlation was found.

2. *Both ordinal regression and multinomial logistic regression under-perform relative to random forest and Knn in socio-economic classification problem*

   Despite being the only model that takes the ordered nature of socio-economic classes into account, the proportional odds model ultimately performed no better than a biased random guess, assigning households to either C1 or C2. While previous studies have shown links between the amount of energy consumed and disposable income (which is itself associated with socio-economic class) [24], analysis of the HES data as well as other studies on domestic energy consumption found that not just lifestyle, but also dwelling-specific factors contributed to the electricity consumption of households. This is suggestive that, while it is possible to infer a households' socio-economic status, more sophisticated methods need to be used to account for these factors, such as a neural network with deep hidden layers.

6.2.2 Random Forest

1. *Improved performance not guaranteed using SFS*

   Indicated by the results of the multi-class problem, employing statistical methods to reduce the dimensionality of the features used by the random forest classifier do nott give better results on test data.

6.2.3 Knn

1. *The larger K is, the more skewed the trade-off between TPR and FPR becomes*

During the optimisation phase, the number of nearest neighbours was adjusted to find the optimum value. As K increased, the accuracy and TPR continued to increase whereas the TNR declined until performance was equivalent to a biased guess. This was due to the imbalances in class sizes and can be adjusted for by limiting the size of K. Conversely, using small values of K such as 1 or 2 also produced poor results because the classes were not separable. Using the 5 nearest neighbours to evaluate the class of a new point gave the best results in cross-validation.

2. *Euclidean distance gives the most reliable results*

   MATLAB's built-in Knn predictor allows numerous different distance measures to be specified including euclidean, Hamming, Manhatten and Mahalonobis. Of these, euclidean distance was found to give the best results based on TPR and TNR trade-off. Euclidean distance assumes that each dimension is equally weighted which is reasonable since feature selection methods were used to determine a subset of features that reduce noise.

6.3 Comparison to Previous work

Comparing the results obtained here to those found by Beckel et al. [22, 23, 10] who used smart meter data from Irish households to predict the socio-economic class and presence or absence of children, our classifiers performed better. While Beckel et al. were able to construct an SVM classifier that had an accuracy of 71% and MCC of 0.32 in discriminating between households with and without children, the results here showed that it is possible to obtain an accuracy of 83.7% and MCC of 0.64 using logistic regression — a notable improvement. While Beckel et al. achieved similar accuracy in their attempt to predict socio-economic class, the maximum MCC they obtained was 0.19, whereas the random forest we constructed had an MCC of 0.4. While this is not a strong correlation, it does indicate stronger predictive power than previous attempts.

Chapter 7

Conclusion and Further Work

7.1 Conclusion

The communications network for the UK's roll out of a nationwide smart meter grid is, as of this writing, more than 91% complete [52]. As energy suppliers embark on retrofitting smart meters in homes throughout the nation, concerns about what can be inferred from the data being collected, stored and transmitted are being raised [4]. The goal of this project was to see if it is indeed possible to construct models that could successfully infer detailed personal information from smart meter data. Using a variety of supervised learning methods, we have been able to confirm that extracting certain meaningful characteristics about the occupants of a household from smart meter data is feasible.

The main contributions of this project were:

- **Dataset** : Data pertaining to the electricity consumed by households that participated in the HES study was extracted and inserted into a MySQL database. This raw data was manipulated to construct a dataset of 519 labeled time-series, each of which was of uniform length and granularity.

- **Feature Engineering** : The data was extracted from the database and imported to MATLAB where it could be further analysed to search for discrepancies between consumption patterns that could be indicative of different household groups. The specific features constructed were detailed in Chapter 3.

- **Classification** : Two household characteristics that might be of interest to third parties were chosen to focus on for this project: whether or not children were present and Ipsos MORI socio-economic class, and classifiers were trained using the features to tackle the classification problems at hand.

  For the classification of children present in a household, 6 models were constructed: two logistic regression models, two random forests and two Knn classifiers. Testing the classifiers on unseen data showed that predicting whether or not children are present in a household is possible with an accuracy of 83% and MCC of 0.65.

  Meanwhile 8 models were constructed to try to predict the socio-economic class of a household from its electricity consumption. This was found to be

a more difficult task than that of inferring the presence of children. The best results were generated by a random forest, with accuracy of 57% and MCC of 0.41. While this is 1.5 times better than the baseline accuracy, the confusion matrices would indicate that it is still largely dependent on the bias in the sample population.

Nonetheless, the results show that models can be constructed to predict household characteristics using electricity readings such as the kinds that will be sent to energy suppliers in the years ahead. In addressing the issue of what privacy intrusions one can reasonably expected, it is important to keep in mind that some characteristics are easier to identify than others.

7.2 Further Work

Possible further research and areas of improvement include:

- Using the HES dataset (and/or similar surveys in future) to predict more properties of a household and dwelling. The socio-economic and presence-of-children problems were chosen because, of the questions answered in the HES questionnaire, and research into the kinds of data sought by energy suppliers and third parties, they were assumed to be of interest to someone wishing to know more about the inhabitants. Other information was also gathered, such as the number of occupants, their employment status, and views on environmental issues. Dwelling specific information was captured as well, such as the age of the property and the number of square feet. Models similar to those presented here could be created to predict these and other sorts of characteristics.

- More sophisticated models, such as neural networks, could be created that factor out the dwelling-specific influences or other latent factors, such as the number of occupants. Alternatively, *a priori* knowledge could be assumed and used as features to boost performance.

- The UK government has already considered the issue of granularity and concluded that the smart meter information will be transmitted to the customers in near real time, but to energy companies in 30-minute intervals [2]. Therefore, it would be worthwhile to look specifically at what information can be extracted from half-hourly consumption data, readings such as was done by Beckel et al. and McLoughlin et al.[23, 24]

- The only ordinal classifier used in performing socio-economic classification was ordinal logistic regression. Methods, such as those introduced by Eibe Frank and Mark Hall [53] allow an otherwise nominal model to treat classes as ordinal without modifying the underlying learning scheme. This could be exploited to train random forests and Knn methods to identify a household's socio-economic group, as well as other household properties, such as the number of occupants or their views on environmental issues (which are known from the HES questionnaire).

Bibliography

[1] Office for National Statiscics. *Full Report: Household Energy Spending in the UK, 2002-2012*. 2014.

[2] Department of Energy and Climate Change. *Smart Metering Implementation Programme - Data access and privacy: Government response to consultation*. 2012.

[3] Stop Smart Meters! (UK). Stop smart meters! (uk), 2015.

[4] Ross Anderson. Smart meter security: a survey. Technical report, Cambridge University, 2011.

[5] SAS United Kingdom. *How SAS can help energy companies take advantage of the data explosion*. From smart metering to smart marketing. 2010.

[6] Energy Saving Trust. *Smart Homes Integrating Meters Money Energy Research*. Energy Saving Trust, 2011.

[7] Department of Energy and Climate Change. *Smart Metering Implementation Programme - Consultation on the second version of the Smart Metering Equiptment Technical Specifications*. 2012.

[8] Department of Energy and Climate Change. *The Smart Metering System*. 2012.

[9] Department of Energy and Climate Change. *Smart Metering Implementation Programme - Communications Hub Technical Specifications*. 2012.

[10] Christian Beckel, Leyna Sadamori, Thorsten Staake, and Silvia Santini. Revealing household characteristics from smart meter data. *Energy*, 78:397–410, 2014.

[11] Gianfranco Chicco. Overview and performance assessment of the clustering methods for electrical load pattern grouping. *Energy*, 42(1):68–80, 2012.

[12] Hong-An Cao. *Are domestic load profiles stable over time? An attempt to identify target households for demand side management campaigns*, pages 4733 – 4738. IEEE, 2013.

[13] J. Z. Kolter and Tommi Jaakkola. Approximate inference in additive factorial hmms with application to energy disaggregation. In Neil D. Lawrence and Mark A. Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS-12)*, volume 22, pages 1472–1482, 2012.

[14] Mikhail A. Lisovich, Deirdre K. Mulligan, and Stephen B. Wicker. Inferring personal information from demand-response systems. *IEEE Security and Privacy Magazine*, 8(1):11–20, 2010.

[15] Harold Wilhite, Hidetoshi Nakagami, Takashi Masuda, Yukiko Yamaga, and Hiroshi Haneda. A cross-cultural analysis of household energy use behaviour in japan and norway. *Energy Policy*, 24(9):795–803, 1996.

[16] Intertek. *Household Electricity Survey A study of domestic electrical product usage.* 2012.

[17] Energy Saving Trust. Powering the nation: Household electricity-using habits revealed.

[18] Nicola Terry Jason Palmer. Powering the nation: Electricity used in homes and how to reduce it.

[19] Jason Palmer, Nicola Terry, and Tom Kane. *Early Findings: Demand side management.* 2013.

[20] Household electricity survey: Cleaning the data.

[21] Irish Social Science Data Archive. Issda — commission for energy regulation (cer), 2015.

[22] Christian Beckel, Leyna Sadamori, and Silvia Santini. Automatic socio-economic classification of households using electricity consumption data. *Proceedings of the fourth international conference on Future energy systems*, pages 75–86, 2013.

[23] Christian Beckel, Leyna Sadamori, and Silvia Santini. *Towards automatic classifi-cation of private households using electricity consumption data*, pages 75–86. ACM, 2013.

[24] Fintan McLoughlin, Aidan Duffy, and Michael Conlon. Evaluation of time series techniques to characterise domestic electricity demand. *Energy*, 50:120–130, 2013.

[25] Infuse2011.mimas.ac.uk. Infuse data wizard, 2015.

[26] Cse.org.uk. How much electricity am i using? — centre for sustainable energy, 2015.

[27] Department of Energy and Climate Change. *Domestic energy use study: to understand why comparable households use different amounts of energy.* 2012.

[28] Fabian Moerchen. *Time series feature extraction for data mining using DWT and DFT.* 2003.

[29] Jason Osborne. Notes on the use of data transformations. *Practical Assessment, Research and Evaluation*, 2002.

[30] Morgan C. Wang and Brad J. Bushman. Using the normal quantile plot to explore meta-analytic data sets. *Psychological Methods*, 3(1):46–54, 1998.

[31] Leticia M. Blazquez Gomez, Massimo Filippini, and Fabian Heimsch. Regional impact of changes in disposable income on spanish electricity demand: A spatial econometric analysis. *Energy Economics*, 40:S58–S66, 2013.

[32] M. Bartley and C. Owen. Relation between socioeconomic status, employment, and health during economic change, 1973-93. *BMJ*, 313(7055):445–449, 1996.

[33] Teachingintheuk.com. Teaching jobs — supply teaching jobs - teaching personnel, 2015.

[34] Careerbuilder.co.uk. Find jobs on careerbuilder.com, 2015.

[35] Mark S Nixon and Alberto S Aguado. *Feature extraction and image processing for computer vision*. Academic Press, 2012.

[36] Amara Graps. An instroduction to wavlents. *IEEE Computational Science and Engineering*, 2(2):50–61, 1995.

[37] Mark A. Hall and Lloyd A. Smith. Feature selection for machine learning: Comparing a correlation-based filter approach to the wrapper, 1999.

[38] Andre Elisseeff Isabelle Guyon. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 2003.

[39] Ruihu Wang. Adaboost for feature selection, classification and its relation with svm, a review. *Physics Procedia*, 25:800–807, 2012.

[40] Somol P., P. Pundil, J. Novicova, and P Pacli'k. Adaptive floating search methods in feature selection. *Pattern Recognition Letters*, 20(11-13):1157–1163, 1999.

[41] Juha Reunanen. Overfitting in making comparisons between variable selection methods. *Journal of Machine Learning Research*, 2003.

[42] Carl Edward Rasmussen and Christopher K. I Williams. *Gaussian processes for machine learning*. MIT Press, 2006.

[43] Peter McCullagh. Regressuib models for ordinal data. *Journal of the Royal Statistical Society*, 1980.

[44] Klaus Ob ermayer Ralf Herbrich, Thore Graep el. Regression mo dels for ordinal data: A machine learning approach. Technical report, Technical University of Berlin, 1999.

[45] leo Breiman. Random forests. *Machine learning*, 2001.

[46] Leif E. Peterson. K-nearest neighbor, 2009.

[47] JERZY STEFANOWSKI. Data mining - evaluation of classifiers. Poznan University of Technology.

[48] Lisa Gaudette and Nathalie Japkowicz. title = Evaluation Methods for Ordinal Classification,. In *Advances in Artificial Intelligence*.

[49] Willem Waegeman, Bernard De Baets, and Luc Boullart. Roc analysis in ordinal regression learning. *Pattern Recognition Letters*, 29(1):1–9, 2008.

[50] David M W Powers. Evaluation: From precision, recall and f-factor to roc, informedness, markedness and correlation. Technical report, University of South Australia, 2007.

[51] J. Gorodkin. Comparing two k-category assignments by a k-category correlation coefficient. *Computational Biology and Chemistry*, 28(5-6):367–374, 2004.

[52] Smartdcc.co.uk. Data communications company, 2015.

[53] Eibe Frank and Mark Hal. A simple approach to ordinal classification. Technical report, University of Waikato, 2001.

Appendix A

Children

SFS			Manual
Log Reg	**KNN**	**Rand Forest**	
Mon Daytime	Mon Evening	Thur Total	Month Total
Tue Evening	Mon Night	Sun Night	Sun Daytime
Wedy Night	Wed Evening	Mon Morning	Sat Evening
Thur Daytime	Fri Morning	Mon Daytime	Thurs Variance
Fri Morning	Sat/Weekday Ratio	Mon Evening	Mon Morning
Tue Variance	Sun Variance	Tue Daytime	Fri Evening
Thury Variance	Mon Variance	Fri Night	Sat Variance
Saty Variance	ρ(Mon Thur)	Thur Variance	Mon Total
ρ(Mon Tue)	ρ(Mon Friy)	Fri Variance	Sat Total
ρ(Wed Thur)	ρ(Tue Wed)	ρ(Mon Wed)	ρ(Mon Tue)

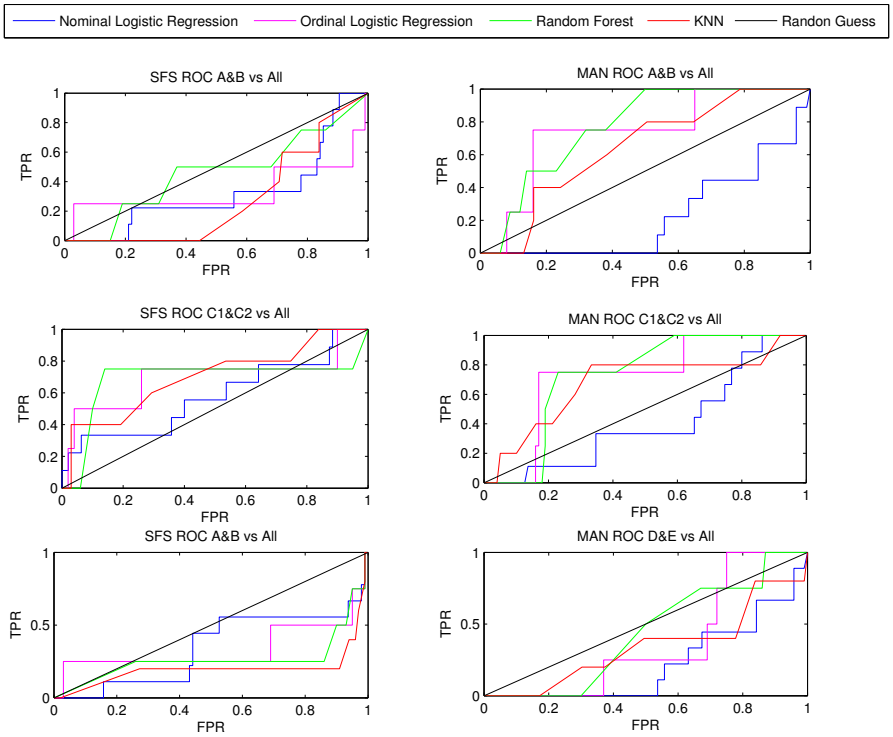
Table A.1

Social Grade

SFS				Manual
Ord Log Reg	**Nom Log Reg**	**KNN**	**Rand Forest**	
Mon Morning	Fri Total	Fri Total	Sun Total	Mon Night
Tue Morning	Mon Morning	Tue Morning	Wed Total	Tue Variance
Tue Daytime	Tue Morning	Wed Night	Mon Morning	Mon Total
Fri Night	Tue Daytime	Thur Morning	Mon Night	Tue day
Fri Variance	Wed Morning	Fri Morning	Wed Morning	Sun Night
Sat Variance	Wed Daytime	Sat Evening	Wed Evening	Thur Night
ρ(Mon Tue)	Wed Evening	Sun Total	Fri Night	ρ(Tue,Fri)
ρ(Mon Fri)	Friday Morning	Thur Variance	Thur Variance	Thur Total
ρ(Wed Fri)	Fri Night	Fri Variance	ρ(Mon Fri)	Sat Night
First Fourier Feature	Sat Daytime	ρ(Wed, Thur)	First Fourier Feature	Fri Evening

Table A.2

Figure A.1: ROC curves of socio-economic classifiers



Nominal Logistic Regression Weights

	boundary 1	**boundary 2**	**boundary 3**	**boundary 4**	**boundary 5**
w_0	51.3976	19.566	43.2567	36.3944	38.2503
w_1	3.8958	2.7275	6.027	3.3391	4.5195
w_2	-2.7457	0.1167	-2.6408	-1.4902	-1.2886
w_3	4.1992	3.3548	2.9044	1.8177	2.7361
w_4	-1.7274	-2.4818	-2.6266	-3.7323	-3.3666
w_5	-0.2175	-2.7121	-1.8633	-0.0991	-1.088
w_6	3.0446	1.175	2.9184	1.8555	1.4066
w_7	-2.2498	-0.8725	-3.4686	-2.1848	-2.873
w_8	-4.0052	-1.4543	-2.2931	-1.7864	-2.1979
w_9	-3.437	-0.5247	-1.1193	-1.2472	0.1815
w_{10}	0.0349	-0.0294	0.3932	1.6358	0.2633