Power to The People: Analysing and Predicting Household Power Consumption

Trihao Van

6th November 2022

Contents

1 Introduction

As of 2020, the energy supply sector is the second highest emitter of greenhouse gasses in the UK, accounting for 21% of total greenhouse gas emissions[1].

With the price of energy continuing to rise, and increasing gravity being placed on climate and sustainability, **How might we help consumers reduce power consumption and make power supply more efficient?**

This project aims to answer this question using smart meter data from households in London on fixed and time of use (variable) tariffs. Households on a variable tariff are notified of changes in energy price via text message and smart meter notifications. This tariff hypothetically encourages customers to modify their energy usage, potentially evening out demand, and reducing consumption. Households on a fixed tariff are receive the same cost of energy throughout the day.

From this data we will determine whether switching to a variable energy tariff is an effective way to reduce household energy consumption. A selection of machine learning models will then be used to forecast the data. This could be used in industry to better synchronise supply with demand and thus improve grid efficiency.

2 Exploratory Data Analysis

2.1 Loading and Cleaning

The dataset contains half hourly power readings from smart meters in 5,567 London households between November 2011 and February 2014[1], totalling $\tilde{1}68$ million rows of data.

115,453 duplicate rows, which made up 0.7% of the data, were dropped. These were deduced to be erroneous entries, as each household should only have a single reading for each time interval. 5,560 null values which were retained, these would be ignored during the following aggregation process.

There were notably a significant number of zero readings, which induced abnormally large variances in the data. These were also dropped as it is unlikely that any household, even when unoccupied, would not be consuming any power at all. It is possible that these readings, many of which occurred at the start of the data, resulted from smart meters which were malfunctioning or simply not switched on.

In order to make the dataset more manageable given the available computing resources, the data was aggregated by tariff into hourly, daily, and weekly periods. This significantly reduced the number of rows that needed to be processed in successive operations.

2.2 Plotting

From the daily plot in Figure 1, we can immediately see that there is a significant difference in power consumption between tariffs. Households on a variable tariff consistently consume $\tilde{1}0\%$ less power than those on a fixed. Interestingly, the pattern of use remains roughly identical - The same observation was also made in the weekly and hourly plots. This is likely due to the fact that not all variable tariff households respond to price changes simultaneously, perhaps due to being out of the home or asleep. Price changes also do not occur at the same time each day, limiting the possibility of preemptive changes in energy use.

Strangely, we can also notice several abnormally low values at the start of the data. This again, could be attributed to malfunctions during the initial setup of the meters. For subsequent steps in the project, only data from the start of 2012 would be considered to eliminate the effects of these readings.

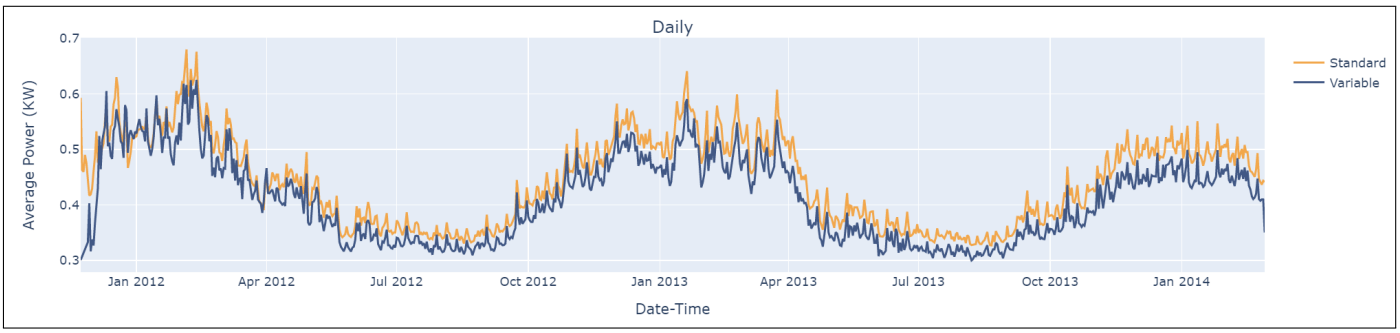

Figure 1: Average daily power consumption

We can see several levels of seasonality in the data. As we would expect, consumption is significantly higher during the winter months than in the summer. We also see a day to day seasonality. Here the peaks are on the weekends, when people spend more time in their homes. Further seasonality still can be seen in Figure 2 where there is a peak in consumption around 7pm, when people come home from work, and a trough overnight. There is also a slight increase in the difference in consumption around 12am. This could be due to variable tariff households preemptively switching devices off in case unactionable price changes occur overnight.

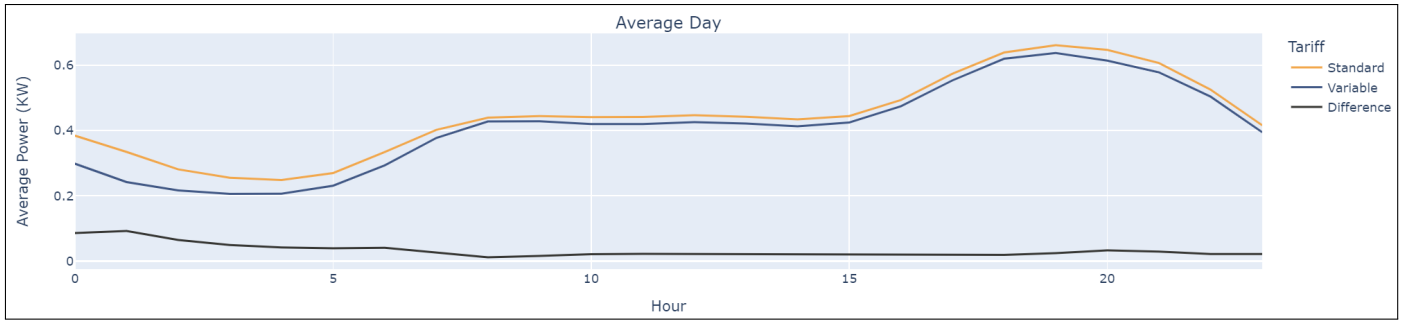

Figure 2: Hourly power consumption on an average day

3 Modelling

In this section we will explore four different machine learning models forecasting at various ranges and granularities:

1. Linear regression - Short range forecasts, any granularity
2. SARIMA - Mid-long range forecasts, weekly granularity
3. Exponential Smoothing - Mid-long range forecasts, weekly granularity
4. Facebook Prophet - Mid-long range forecasts, daily granularity

Models will be evaluated using mean absolute percentage error (MAPE), as this metric scales with the magnitude of the data. Where possible, this will be include a cross validated average.

3.1 Linear Regression - Baseline Model

The linear regression models aim to predict a single following period of power consumption for all combinations of tariff and granularity, resulting in a total of six models. Whilst these models can make only short range forecasts, they can be useful in providing real time power usage predictions to end users, informing on how current usage behaviour may affect immediate future consumption.

Nine features - lag, rolling average, and rolling variance, across three time intervals - were considered. A cross validated grid search was then carried out on the estimators below to find an optimised model for each dataset.

Scaler:

- Standard scaler - Set mean to 0 and standard deviation to 1
- None

Dimension Reducer:

- Principal component analysis (PCA) - Linearly transform features to maximise variance capture with fewer dimensions
- SelectKBest - Select the single most correlated feature
- None

Regulariser:

- Lasso, L_1 - Cost function scales with the magnitude of the feature coefficient, $|\beta|$
- Ridge, L_2 - Cost n scales with the square of the feature coefficient β^2
- None

Tariff	Granularity	Scaler	Dimension Reducer	Regulariser	Cross Validated MAPE	Test MAPE
Standard	**Hourly**	None	PCA (5 components)	None	5.45%	5.44%
	Daily	None	PCA (6 components)	None	3.73%	2.85%
	Weekly	Standard	PCA (8 components)	None	3.39%	2.71%
Variable	**Hourly**	Standard	PCA (6 Components)	None	6.62%	6.33%
	Daily	None	PCA (6 components)	None	4.00%	2.97%
	Weekly	None	PCA (8 components)	None	3.37%	2.45%

Table 1: Grid search optimised linear regression models

The resulting models and their performance can be seen in Table 1. Here we can see that the models all perform generally well. This is likely due to the relatively small variance from point to point, which would also explain the improved performance at larger granularities where the aggregation has further reduced the variance.

3.2 SARIMA

A **S**easonal **A**uto-**R**egressive **I**ntegrated **M**oving **A**verage (SARIMA) model considers autoregressive (lagged), integrated (differenced), and moving average values of a dataset when making predictions. The model also takes into consideration the same values for the seasonal component of the data, as well as the seasonal period.

This, along with the subsequent models in this report, have the ability to forecast much further ahead than the linear regression models.

A weakness of the SARIMA model is the inability to explicitly process multiple seasonalities. Naturally this could be problematic when training on daily or hourly data. For this reason, as well as limited computational resources, we will use only the weekly data.

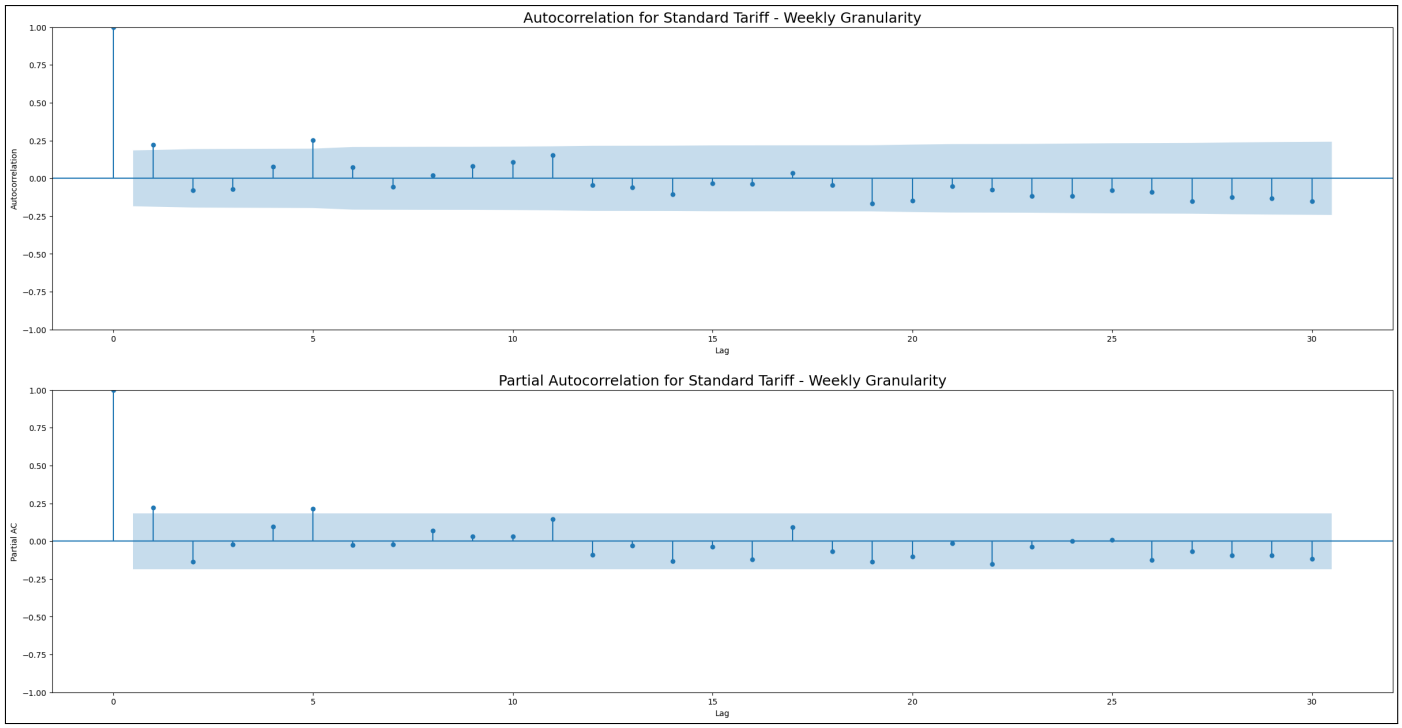

Figure 3: ACF and PACF plots for weekly standard tariff data

Model parameters were informed using auto correlation function (ACF), and partial auto correlation function (PACF) plots, as shown in Figure 3 above. Points outside the blue area are significant, and should be considered in the model. Further tuning was informed by the significance, P values in the summary of each model iteration. The performance of the tuned models can be seen in Table 2 below.

Tariff	**Cross Validated MAPE**	**Final Fold Train MAPE**	**Final Fold Test MAPE**
Standard	5.45%	2.13%	5.44%
Variable	6.62%	2.27%	6.33%

Table 2: Tuned SARIMA model errors

3.3 Exponential Smoothing

Exponential smoothing is a time series method that considers a window of past data points, assigning exponentially reduced weights to points further away to that being predicted. A simple exponential smoothing algorithm looks like[2]:

$$\ell_0 = y_0$$
$$\hat{y}_{t+h|t} = \ell_t = \alpha y_t + (1-\alpha)\ell_{t-1}$$
$$t > 0, \quad 0 < \alpha < 1$$

Where $\hat{y}_{t+h|t}$ is the predicted point, t is time, y is the actual data point, and α is a smoothing factor.

Expanding to ℓ_{t-2}, and substituting ℓ_{t-1} we can see the exponential nature of this formula.

$$\ell_{t-1} = \alpha y_{t-1} + (1-\alpha)\ell_{t-2}$$
$$\ell_t = \alpha y_t + (1-\alpha)y_{t-1} + (1-\alpha)^2\ell_{t-2}$$

We can see that as the number of terms increases, the weighing progresses geometrically by factors of $(1-\alpha)$.

The Holt-Winters method, otherwise known as triple exponential smoothing, is a modification of simple exponential smoothing. As the name suggests, two additional formulae are considered, one accounting for a trend in the data, and another for the seasonality.

For a multiplicative model:

$$\ell_t = \alpha\frac{y_t}{s_{t-m}} + (1-\alpha)(\ell_{t-1} + b_{t-1})$$
$$b_t = \beta(\ell_t - \ell_{t-1}) + (1-\beta)b_{t-1}$$
$$s_t = \gamma\frac{y_t}{\ell_{t-1} + b_{t-1}}$$

Where b is the trend component, s is the seasonal component, β and γ are their respective smoothing factors, and m is the seasonal period.

The multiplicative forecast equation combines the smoothing equations:

$$\hat{y}_{t+h|t} = (\ell_t + hb_t)s_{t+1+(h-1) \mod m}$$

Cross validation was not possible in this model due to the limited number of seasons in the data. A rough comparison was made with other models using the train and test MAPEs shown in Table 3.

Tariff	**Train MAPE**	**Test MAPE**
Standard	2.75%	6.60%
Variable	2.56%	3.46%

Table 3: Holt-Winters model errors

3.4 Facebook Prophet

The Facebook Prophet model has a number of advantages over the previously explored models. It can take into account factors such as:

- Multiple seasonalities
- Trend changes
- Holidays
- Exogenous variables

Most relevant to this project is the ability to process multiple seasonalities, the presence of which we have seen in the EDA section. This permits forecasting at a much finer granularity than the SARIMA and Holt-Winters models. For this model we will use the daily data, creating a model which could be used in industry to forecast daily household power demand at medium to long ranges.

The disadvantage of this model however, is the lack of interpretability and limited documentation. The model provides no information on feature contribution and significance which in turn limits the amount of tuning that can be done.

Tariff	**Cross Validated MAPE**	**Best Train MAPE**	**Best Test MAPE**
Standard	8.73%	3.78%	4.53%
Variable	13.52%	3.87%	4.15%

Table 4: FB prophet model errors

4 Conclusions

From the EDA we saw that households on a variable tariff consistently consumed less power than those on a standard tariff despite the pattern of use remaining similar. We can conclude that variable tariffs and notifications of energy price fluctuations are effective ways of making households more aware of their energy use. This in turn effectively encourages a reduction in consumption.

From the MAPEs of the four models, we observe that predicting at longer ranges and finer granularities tends to come at the cost of increased error. However, this could potentially be mitigated by further tuning and data collection. The data used only contained two seasonal cycles which may not have been sufficient for the models to pick up the trend and seasonality during training.

5 Further Work

This project considered only univariate data. Implementing a Prophet or SARIMAX (SARIMA, e**X**ogenous) model with weather data would potentially result in significant improvements in accuracy. Intuitively, and from the EDA, we can assume that factors which encourage people to stay at home, such as cold weather and rainfall, correlate with an increase in power consumption.

Computational and time restrictions were also a limiting factor in this project. Using cloud computing resources to train models on a larger dataset which captures more than two seasonal cycles, along with the aforementioned exogenous data would likely further reduce prediction errors.

References

[1] Office of National Statistics. *2020 UK Greenhouse Gas Emissions, Final Figures*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1051408/2020-final-greenhouse-gas-emissions-statistical-release.pdf. Accessed: 2022-11-06. 2020.

[2] Rob J Hyndman and George Athanasopoulos. *Forecasting: principles and practice, 2nd edition*. Melbourne, Australia: OTexts, 2018.