Optimisation of Behind the Meter DER Generation Assets within Network Constraints: A Roadmap to Successful DR Program (Project 69)

Work Package-1 Final Report

RMIT University, Melbourne

September 2022

Project Lead:	A/Prof. Lasantha Meegahapola
Research Team:	A/Prof. Mahdi Jalili Dr Richardt Wilkinson Dr Kazi Hasan Prof. Brendan McGrath Dist. Prof. Xinghuo Yu Prof. Flora Salim Dr Hui Song Dr Ammar Kamoona Dr Mingchen Gu Mr Kian Keshavarzian Mr Chad Wanninayaka Mudiyanselage Dr Reza Razzaghi Dr Mohsen Khorasany
Prepared for:	C4NET, AGL, AusNet Services
Contact:	A/Prof. Lasantha Meegahapola
Email:	*lasantha.meegahapola@rmit.edu.au*

Acknowledgements: The project team wishes to acknowledge the data support and the feedback received from Mr. Kedem Levy and the AGL team. The team also wishes to acknowledge the Gippsland Water Factory, Federation University, and Greater Western Water for sharing their energy data with the Project-69 team.

Executive Summary

Demand response (DR) programs can benefit electricity consumers, distribution network service providers, and system operators. However, the complexity and characteristics associated with the loads connected to customer premises and the baseline calculation process still require considerable research to exploit maximum benefits from DR programs. The challenges of DR programs are not limited to technical aspects since there are several non-technical elements to consider when implementing DR programs, such as regulatory policy compliance. C4NET launched this DR research program to use the research expertise of member universities to solve some of the industry challenges concerning the DR schemes and accelerate their deployment in electricity networks. This project is expected to produce potential solutions for some of the DR challenges associated with commercial and industrial (C&I) customers. It is envisioned to propose machine learning-based calculation methods to improve upon the baseline calculation process. The project consists of three work packages, and this project report outlines the summary of work completed under work package 1 (WP1).

Work package 1 (WP1) aims at using interpretable machine learning (ML) tools to assess DR programs in C&I customers. This study analysed data from several C&I customers from telecommunications companies, water utilities, university, chemical plant, and shopping centre businesses. An essential part of DR programs for C&I customers is the baseline used to calculate the monetary benefits to customers. The study also investigated how temperature combined with consumption history and simple and explainable machine learning methods may impact the customers' benefit by providing more accurate demand forecasts.

Study has found that machine learning methods can provide more accurate demand predictions than the traditional baseline method (i.e., average baseline method or Baseline Type I), which benefit C&I customers during DR events. Each C&I customer portfolio has a different demand profile and has different correlations with weather parameters; therefore, recommendations have been made for each portfolio separately. The strength of the dataset (i.e., resolution and duration of the dataset) has a strong influence on the performance of the ML technique applied for demand prediction.

Moreover, this study has drawn following specific outcomes and recommendations for each C&I customer group:

- The ML model, namely linear regression (LR) and nonlinear regression (NLR), performs better than the average baseline method with a varying level of significance depending on the customer profiles. Thus, this study recommends using LR and NLR for the baseline calculation for the following C&I customers: chemical plants, telecommunication industry, and water utilities.
- Since LR is much easier to understand than NLR, this study recommends using LR as the baseline model where possible (e.g., water utilities), in which the temperature information can enhance the prediction accuracy of the future energy demand.
- For other C&I customers (e.g., medium manufacturing, metal recycling, sandstone quarry) this study recommends using the average baseline method, since the performance difference is not significant.

- By considering the temperature information, ML models can make more accurate predictions, in particular for water utilities, and shopping centres, as the temperature influences the demand.
- Demand response analysis further shows that more accurate demand predictions provide more potential monetary benefits to C&I customers.

Although the ML models provide the advantages over traditional average baseline method, they require sufficient data for the training process. Therefore, for successful implementation of ML models in DR programs require sufficient data to develop models which can make accurate demand predictions, and hence they can deliver more benefits to C&I customers in DR programs.

Table of Contents

List of Figures

List of Tables

Acronyms and Terminologies

- **AEMO:** Australian Energy Market Operator
- **AER:** Australian Energy Regulator
- **AI:** Artificial Intelligence
- **BLR:** Bayesian Linear Regression
- **CDD:** Cooling Degree Days
- **CNN:** Convolutional Neural Networks
- **C&I:** Commercial and Industrial
- **DER:** Distributed Energy Resources
- **DNSP:** Distributed Network Service Provider
- **DNN:** Deep Neural Networks
- **DR:** Demand Response
- **HDD:** Heating Degree Days
- **LR:** Linear Regression
- **LSTM:** Long Short-Term Memory
- **MAE:** Mean Absolute Error
- **MBL**: Maximum Base Load
- **ML:** Machine Learning
- **NER:** National Electricity Rules
- **NLR:** Nonlinear Regression
- **NMI:** National Meter Identifier
- **RMSE:** Root Mean Square Error
- **SVM:** Support Vector Machines
- **SVR:** Support Vector Regression

- **Baseline:** A baseline is an estimate of the electricity that would have been consumed by a customer in the absence of a demand response event**.**

- **MAE:** This measures the average magnitude of the errors in a set of predictions. It is the average over the samples of the absolute differences between predictions and actual observations.

- **RMSE:** A quadratic scoring rule that also measures the average magnitude of the error. It is the square root of the average of squared differences between predictions and actual observations.

1. Project Overview

RMIT was engaged by C4NET to provide a roadmap for successful demand response (DR) programs for commercial and industrial (C&I) customers. In response to the C4NET request and in collaboration with AGL, the RMIT team proposed a project with three work packages, which are as follows:

- **Work Package-1:** Machine learning for C&I customers' baseline improvement,
- **Work Package-2:** Unlocking the potential of participation of backup generators in DR,
- **Work Package-3:** Identify tariffs that can incentivise the uptake of batteries.

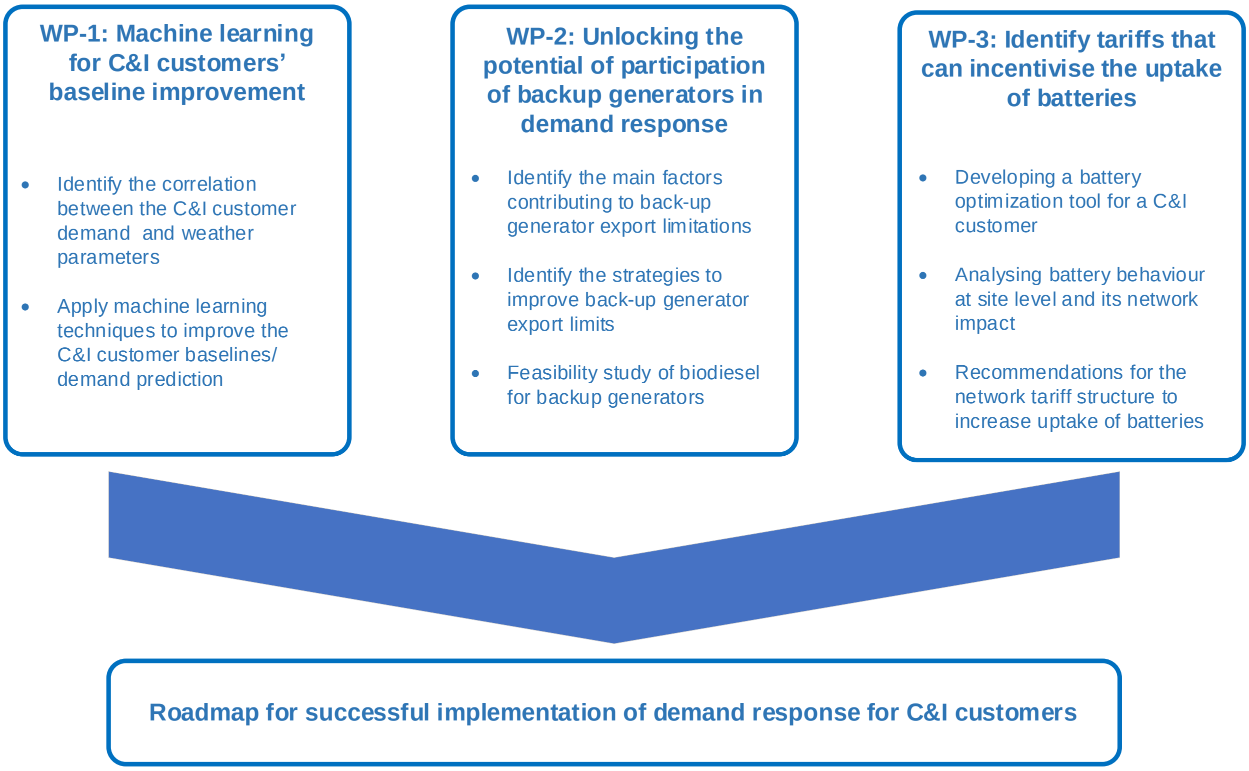

WP-1: Machine learning for C&I customers' baseline improvement
• Identify the correlation between the C&I customer demand and weather parameters
• Apply machine learning techniques to improve the C&I customer baselines/ demand prediction
WP-2: Unlocking the potential of participation of backup generators in demand response
• Identify the main factors contributing to back-up generator export limitations
• Identify the strategies to improve back-up generator export limits
• Feasibility study of biodiesel for backup generators
WP-3: Identify tariffs that can incentivise the uptake of batteries
• Developing a battery optimization tool for a C&I customer
• Analysing battery behaviour at site level and its network impact
• Recommendations for the network tariff structure to increase uptake of batteries
Roadmap for successful implementation of demand response for C&I customers

Figure 1. Aims and objectives of project work-packages.

The WP-1 will apply the machine learning techniques to unlock the potential of demand response schemes, in particular focus on improving the baselines used in DR. WP-2 will focus on unlocking the potential of using backup generators and other embedded generators of commercial and industrial customers in DR programs. It will also explore possible solutions to overcome the regulatory and technical barriers to using backup generators in DR programs. Lastly, WP-3 will explore the role of network tariffs in DR programs which can incentivise the uptake of batteries.

Furthermore, the project team expects to engage the distributed network service providers (DNSPs) to better understand the technical and regulatory barriers to deploying backup generators for commercial and industrial customers in DR programs such as water utilities.

Finally, this project aims to create a roadmap for successful implementation of DR for C&I customers. As a direct outcome of this project, it is envisaged that more commercial and industrial customers will sign up for the DR programs in the future.

2. Scope of Work Package-1

Work Package-1 (WP1) aims to use Machine Learning (ML) tools to assess the concurrent inclusion of temperature information and consumption profiles in improving baselines used to calculate customer monetary benefits. In brief, this study has undertaken the following research activities:

1. Data preparation and pre-processing to feed numerical analysis models;
2. Investigating the baselines and adjusted baselines used in the market;
3. Analysis of historical meter data to qualitatively assess the potential of using an alternative baseline logic for a demand response program to impact market price outcomes;
4. Calculating the key average factors for each customer to provide a better general overview for each customer;
5. Considering metadata such as temperature, high- and low degree days (HDD and CDD), heat- and cold wave duration, weekends and holidays to investigate their impact on improving baselines;
6. Categorising portfolios based on their key factors and providing insight by comparing the customer segmentation for each factor;
7. Assessing baselines based on explainable machine learning models and comparing them with the conventional baselines used in the market.

The project team used three primary data sources for the study;

- AGL dataset comprised of ten (10) C&I customer portfolios
- Gippsland Water Factory dataset
- Greater Western Water dataset

The application of ML techniques for each dataset is explained in the subsequent sections of the report.

3. Determination of Baseline for Demand Response

3.1. Application of machine learning techniques to the AGL dataset

We have been provided with load profiles (consumption data recorded from smart meters) for several C&I customers. In most cases, the recordings were performed in 30-minute time intervals, but for some days (in the case of three customers), the dataset includes 5-minute interval recordings. All the data was considered in the analytics performed for this project.

3.2. Initial data analytics results

The data was provided in Microsoft Excel spreadsheets with attributes, such as event date and time, total power consumption, national meter identifier (NMI), curtailment, baseline, and adjusted baseline. Data pre-processing was required to address inconsistencies and omissions in the dataset, and this was performed using Microsoft Excel. The customers can be categorised into 12 portfolios as follows:

- Chemical Plant
- Medium Manufacturing
- Metal Recycling
- Sandstone Quarry
- Shopping Centre
- Telecom
- Telecom VIC
- University
- Water Utility 1
- Water Utility 2
- Water Utility 5
- Water Utility 6 VIC

The time range variation for each business lies in the interval from 22-12-2019 to 22-07-2021. A comparison between portfolios with regard to Average Power Consumption (MW) is provided in Table 1.

Table 1. A comparison between portfolios in terms of load and energy consumption.

Customer Portfolio	Average MW
Chemical Plant	9.7
Medium Manufacturing	0.5
Metal Recycling	2.6
Sandstone Quarry	1.7
Shopping Centre	9.0
Telecom	15.9
Telecom VIC	13.6
University	2.8
Water Utility 1	0.7
Water Utility 2	7.2
Water Utility 5	0.3
Water Utility 6 VIC	5.3

3.3. Machine learning and data analytics for better assessment of DR programs

Machine learning (ML) is a rapidly growing field, which brings together the fields of mathematics, engineering, and computer science, and may be thought of as creating a computer-generated inference based on past observations and future expected parameters. Over time, as the model encounters more examples of a behaviour, its predictions become more accurate, so that more benefit can be achieved for DR programs using ML [1], [2]. Also, some baseline models only use the same historical time stamps to calculate the value at the same future time (average value), e.g., using the 10:00 historical data to predict 10:00 in the future with a time window (the number of historical days used as inputs), which does not consider the continuity and dependency of the time series demand data or exogenous factors, such as temperature and humidity. In this regard, ML models, including traditional ML models, such as linear regression (LR), nonlinear regression (NLR), and support vector machines (SVM), and deep neural networks (DNN) models, such as long short-term memory (LSTM) and convolutional neural networks (CNN), cannot only take different categories of information as inputs and learns the relationship between the inputs and outputs (demand in this report), but also can interpret how and why these inputs influence the outputs [3], so that more accurate prediction can be achieved.

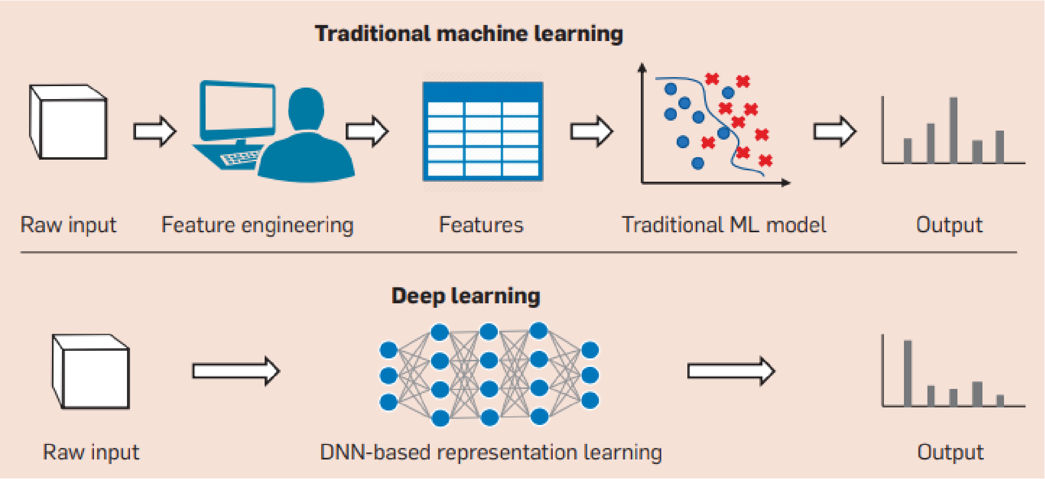


Figure 2. A traditional ML pipeline using feature engineering, and a ML pipeline using DNN-based representation learning [3].

The pipelines of traditional ML and DNN-based learning models are shown in Figure 2, in which the working processes of both types are clearly illustrated [3]. Traditional ML models mainly focus on using feature engineering, where the features are generally interpretable, and the role of ML is to map the representation to outputs. Different from traditional ML models, DNN-based models cannot provide the mapping from representation to output, but it learns representation from the raw data. Therefore, this study chooses several traditional interpretable ML approaches for the DR analysis.

DR programs are very imperative for the energy sector and offer many opportunities. The aging of the power generation fleet in Australia combined with the drive to decarbonisation raises the importance of demand response and load flexibility in general. These DR programs require increased customer participation, with the outcomes depending on program performance and measurement. Such developments result in reduced revenue losses, enhanced insight for regulatory bodies and improved load management. Utilities expect cohesive data and innovative analytics to deliver these benefits, but some factors are barriers to their efforts.

Assessing the performance of participation in DR programs is in many cases dependent on analytical calculations and cannot be measured directly. The reason is that there needs to be a comparison between the measurement and what the demand profile would have been if it were not for the event, i.e., the actual figures compared to some assessment of business as usual. This is similar to energy efficiency assessments but on a very time localised basis. This can be challenging when the load profile is not flat and/or dependent on the conditions of the event. Temperature is a good example. DR events are likely to occur on extreme heat days which would be when certain customers use significantly more energy than their typical average. Thus, their demand levels will be higher than the days prior to the event, and so the calculation needs to account for the conditions on the day. Therefore, baselines are used for assessing DR performance.

This is an area that can be supported and further developed by projects such as this one. The current use of baselines is limited and/or may rely on a simple calculation approach rather than a potential best practice. The aim of this research is to assess whether and how ML can be leveraged as a mathematical tool to create a better baseline approach. The efforts to improve DR programs will unavoidably fail unless DR teams can attain a high level of transparency, data consistency and analytic power to overcome these problems. ML and data analytics can potentially be used to overcome some of the above challenges. The following are potential contributions that ML tools may make in DR programs, and only some of them are addressed in this project due to the limitations of data available for this study.

3.4. Alternative (ML-based) baseline calculations

The most important feature in DR programs is the baseline. A baseline is '*an estimate of the electricity that a customer would have consumed in the absence of a demand response event*' [4]. The dataset includes baseline logic to calculate the baseline and adjusted baseline for its DR program. In all DR programs, the baseline or measurement and verification (M&V) plays a crucial role in determining the magnitude of the resource and its contribution to supporting the grid. M&V also defines customer return for their contribution, which will affect the number and types of potential customers joining the DR program. There are currently many approaches for calculating the baseline [5]–[8]. Some are more accurate than others in estimating the baseline [6]. There is always a common rule that applies to all approaches: a good baseline design is defined by three important characteristics: *accuracy, simplicity, and integrity*.

3.4.1. Accuracy

Customers should receive credit for the curtailment they provide with an optimal level of accuracy. Thus, a baseline method should utilise as much available data as possible to estimate the load accurately in the absence of a DR event.

3.4.2. Simplicity

The baseline should be simple for all participants to understand and calculate. Moreover, it needs to be possible to estimate the baseline before or during DR events, so that it can be used to monitor curtailment in real-time.

3.4.3. Integrity

A baseline method should not consist of characteristics that allow participants to manipulate their consumption and mislead the system.

Making a trade-off between these three characteristics is not a straightforward task. For example, a baseline that is not easy to manipulate can be complex for stakeholders to work with. On the other hand, a simple baseline might allow participants to game the system. Therefore, baseline consistency and coherency need to be assessed before implementation to ensure that they provide a stable and reliable curtailment.

We have identified the baseline as the topic that ML and data analytics could address and contribute to this project. This can potentially improve C&I customers' contributions in DR by providing more accurate baselines that truly represent customers' contributions to the DR event. We refer to this method as an ***AI-empowered baseline*** for DR events.

The existing baseline methodology is referred to in this report as Baseline Type I.

3.4.4. Baseline Type I

A baseline performance assessment based on historical interval meter data may also include other parameters, such as weather and calendar data. The Baseline Type I method is the most popular method used in DR programs today. Variations of this method include Averaging, Regression, Rolling Average, and Comparable Day.

Characteristics of Baseline Type I methods are as follows:

- Baseline shape is the average load profile.
- Utilises meter data from each individual site.
- Relies upon historical meter data from days immediately preceding a DR event.
- May use weather and calendar data to inform or adjust the baseline.

Averaging Methods

The most broadly used Baseline Type I approaches are the averaging methods, which create baselines by averaging recent historical load data to shape load approximations for specific time intervals. Averaging methods are often called representative day methods or High X of Y methods.

"*A High X of Y baseline considers the Y most recent days preceding an event and uses the data from the X days with the highest load within those Y days to calculate the baselin*e." [9]

A simple example is a simple average across the Y most recent days.

Maximum Base Load

A maximum base load (MBL) evaluates the demand resource's capability to reduce to a certain level of electricity demand. MBL methods find the maximum energy usage expected from each customer and then set a specific level of electricity usage equal to the maximum level minus the dedicated capacity of the customer. MBL methods are sometimes referred to as "drop to" methods because the customer must drop to a specific usage level during an event. On the contrary, most Baseline Type I methods are referred to as "drop by" because the customer knows the amount of committed capacity they must reduce. However, the usage level is not necessarily constant. The MBL is an example of a static baseline because it remains at one level compared to a Baseline Type I method that generates a dynamic, changing load profile throughout the day. Note that with the MBL baseline, it is entirely possible for a customer to “perform” by doing nothing at all, as long as the load is already at or below the “drop to” level [9].

Meter Before – Meter After

The baseline is calculated using only actual load data from a time interval immediately before an event.

3.4.5. Baseline Type II

Most baselines are created using historical meter data from the individual site of the customer. There are cases where data from individual sites are not available, but data from an aggregating meter or a meter representative of several sites are available. In these cases, the meter data can be used to create a baseline for a group of sites and then a method used to allocate the load to specific sites. For example, consider a group of homogenous sites with similar load behaviour. A Baseline Type II method could analyse a few of the sites to develop an average load estimate per site and then allocate load from the aggregated baseline.

In DR programs with commercial and industrial customers, which is the focus of this report, Baseline Type II methods are not common, because most sites either have or can be cost-effectively equipped with interval meters. The Baseline Type II method is more often used in residential DR programs, where it has been cost-prohibitive to install interval meters at every house. However, as the deployment of residential interval meters increases, the need for Baseline Type II methods will likely also decrease [9].

Generation

The baseline is set as zero and measured compared to usage readings from behind-the-meter emergency backup generators. This type of baseline is only applicable to facilities with on-site generation.

The second part of this discussion is allocated to the baseline logic and our findings of implementing an averaging baseline method.

3.5. Preliminary results

3.5.1. Variability

In statistics, the coefficient of variation is used to measure the variability of a data series. More specifically, in the context of this dataset, it measures the spread of the consumer energy demand (kWh). The variability measure will enable us to describe how much data sets vary in terms of the energy demand (kWh) and allows us to compare it with different datasets (e.g., energy variability between the Chemical Plant and Medium Manufacturing measured in kWh). It is calculated by dividing the standard deviation by the average (mean) and is usually expressed in percentage (note that standard deviation is the average deviation (not the arithmetic average, mean) of a data series from its mean):

$$Variability = Coefficient\ of\ Variation = 100 \times \frac{Standard\ Deviation}{Mean} \qquad (1)$$

Table 2. The variability of portfolios.

Customer Portfolio	Variability (%)
Chemical Plant	19.4
Medium Manufacturing	89.7
Metal Recycling	66.9
Sandstone Quarry	45.7
Shopping Centre	10.6
Telecom	23.7
Telecom VIC	7.8
University	18.8
Water Utility 1	31.2
Water Utility 2	11.4
Water Utility 5	47.7
Water Utility 6 VIC	72.5

In terms of variability, it is possible to split the portfolios listed in Table 2 into three categories:

1. **Low variability - up to 33.3%:**
   - Chemical Plant
   - Telecom
   - Telecom VIC
   - University
   - Water Utility 2
2. **Medium variability - from 33.3% to 66.6%:**
   - Sandstone Quarry
   - Shopping Centre
   - Water Utility 1
3. **High variability - from 66.6% to 100%:**
   - Medium Manufacturing
   - Metal Recycling
   - Water Utility 5
   - Water Utility 6 VIC

Since not all the businesses stay fully operational outside of office hours, this factor needs to be calculated during office hours before the event starts to be comparable for each pair of businesses. Therefore, to calculate *variability* and *average power consumption (MW)*, we have averaged the *power consumption* only from 10:00 to 13:00.

The *variability*, calculated along the *date/time* axis, can be a valuable measure to determine the *baseline* logic. Note that the existing baseline logic is a high X of Y most recent days preceding an event. The higher the variability of the portfolio, the more the dependency of the baseline position on X. For example, for a low variability portfolio, the five days with the highest load within a 10-day period might not significantly differ from six days within a 10-day period. However, it can be a significant difference for a high variability portfolio. We expect the proposed AI-empowered baseline to represent the actual customer contribution during higher load variability accurately.

3.5.2. Average power consumption

The average power consumption calculated for each customer is listed in Table 1 and categorised into three categories: less than 2 MW, 2 to 7 MW, and more than 7 MW. The results are as follows:

1. **Low power consumption - less than 2 MW:**
    - Medium Manufacturing
    - Sandstone Quarry
    - Water Utility 1
    - Water Utility 5
2. **Medium power consumption - between 2 and 7 MW:**
    - Metal Recycling
    - University
    - Shopping Centre
    - Water Utility 6 VIC
3. **High power consumption - more than 7 MW:**
    - Chemical Plant
    - Water Utility 2
    - Telecom
    - Telecom VIC

Figure 3 shows the actual power consumption for Shopping Centre and Telecom separately. It can be observed that there is a sharp change in power consumption for Shopping Centre between 7:00 to 10:00 while, that of Telecom is not changing significantly during this period. Since we need to compare the categories based on the data close to the event period, we excluded the data points before 10:00 from this averaging process.

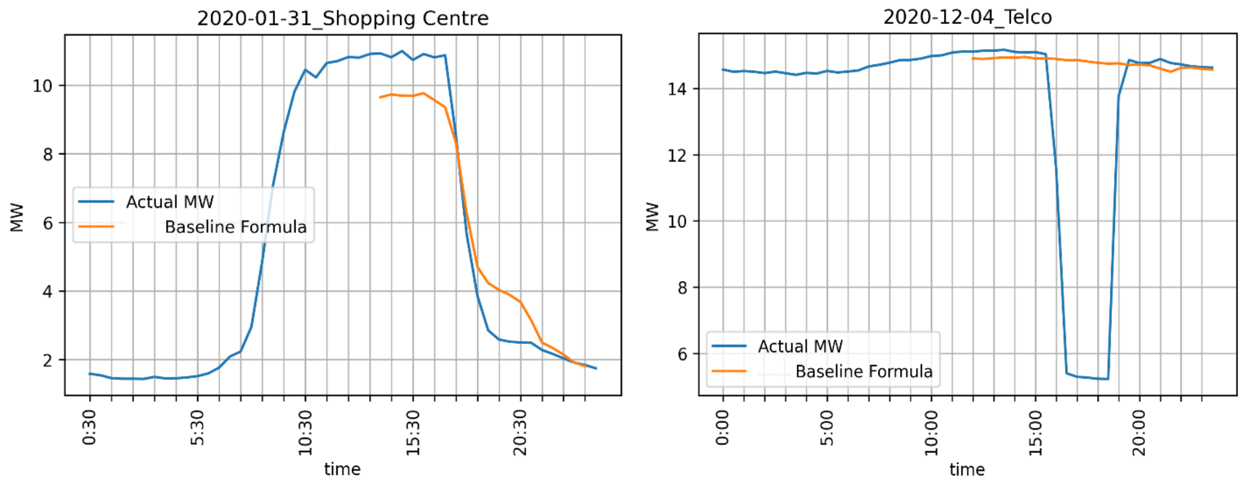


Figure 3. The actual power consumption plotted for a day for: (a) Shopping Centre portfolio (significant power consumption measured from around 10:00); (b) Telecom portfolio (significant power consumption measured from the start of the day).

Actual power consumption (MW) is one of the most important factors for retailers to convince more effective customers to join their DR programs. These consumers can reduce energy consumption significantly during the peak load. The dataset shows that all four high-power consumer portfolios (Chemical Plant, Water Utility 2, Telecom, and Telecom VIC) are also the least variable loads. This means that changing the X and Y (e.g., from top 5 out of 10 business days to middle 5 out of 10 or even 10 out of 10) will have less effect on the baseline position than highly variable portfolios.

3.5.3. Correlation to metadata and other attributes

The most important weather parameter to consider in DR programs is the ambient temperature. This study has used the temperature data from the Australian Government Bureau of Meteorology website for the weather stations closest to Sydney and Melbourne's central business districts (CBDs) [10]. The weather station information is listed in Table 3.

Table 3. The weather station information used to obtain temperature data for Sydney and Melbourne.

Site name	Sydney (Observatory Hill)	Melbourne (Olympic Park)
Site number	066214	086338
Commenced	2017	2013
Latitude	33.86° S	37.83° S
Longitude	151.20° E	144.98° E
Elevation	43 m	8 m
Operational status	Open	Open

The correlation is determined between the *total daily energy consumption* (not listed in Table 1) and the HDD or CDD calculated according to the information provided in Australian Government Bureau of Meteorology [11].

"Heating and Cooling Degree Days (CDDs & HDDs), which indicate the level of comfort, are based on the average daily temperature. The average daily temperature is calculated as follows: [maximum daily temperature + minimum daily temperature] / 2.

If the average daily temperature falls below comfort levels, heating is required and if it is above comfort levels, cooling is required. The HDDs or CDDs are determined by the difference between the average daily temperature and the BASE (comfort level) temperature. The BASE values used are 12 and 18 degrees Celsius for heating and 18 and 24 degrees Celsius for cooling." [11]

As HDD and CDD are categorical data series, we can still use the conventional Pearson correlation coefficient:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (2)$$

where n is the sample size, x_i, y_i are indexed sample points and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean [12]. The results are shown in Table 4.

Table 4. Pearson correlation (%) between the daily total energy consumption and HDD, CDD, weekends and holidays.

Customer	Correlation to HDD (18°C) %	Correlation to CDD (24°C) %	Correlation to Weekends %	Correlation to Holidays %
Chemical Plant	-	0.8	-26.8	-47.4
Medium Manufacturing	-	7.2	-38.2	-28.8
Metal Recycling	-	-4.8	-44.6	-33.4
Sandstone Quarry	-	-18.4	2.4	-25.0
Shopping Centre	-	81.1	-38.9	-15.5
Telecom	45.6	-9.9	5.2	-4.8
Telecom VIC	-16.5	17.0	3.2	6.1
University	-	15.9	-27.1	-58.0
Water Utility 1	-	-6.7	-25.4	-0.7
Water Utility 2	-	11.7	-19.9	-7.0
Water Utility 5	-	-0.3	-44.3	-13.0
Water Utility 6 VIC	-16.0	-	21.5	-

From the above results, the portfolios that can potentially be considered to use temperature as a part of its predictor are Telecom and Shopping Centre. However, in the data preparation stage for the ML models, we will let the model decide how significantly predictors affect the predictability. On the other hand, the Shopping Centre and University portfolios use relatively high energy-consuming cooling and heating devices to keep the inside temperature close to acceptable comfort levels.

Concerning these factors, we recommend using temperature as part of the predictor for the following portfolios listed in Table 1:

1. **Recommended:**
    - Telecom
    - Shopping Centre
2. **May be useful:**
    - Telecom VIC
    - University
3. **Not Recommended:**
    - Chemical Plant
    - Medium Manufacturing
    - Metal Recycling
    - Sandstone Quarry
    - Water Utility 1
    - Water Utility 2
    - Water Utility 5
    - Water Utility 6 VIC

The '*NaN'* (Not a Number) values in Table 4 result from a *non-varying* categorical data series (HDD and CDD) due to the data's limited date span covering only a short hot- or cold period over one year. In addition, only positive correlation values to both HDD and CDD are of interest.

3.5.4. Weekends and Holidays

This study focusses on the high negative correlation to weekends and holidays shown in Table 4. From this point of view, one can categorise the portfolios based on the recommendation of using weekends and holidays as predictors:

1. **Recommended:**
    - University
2. **May be useful:**
    - Chemical Plant
    - Medium Manufacturing
    - Metal Recycling
    - Shopping Centre
    - Water Utility 5
3. **Not Recommended:**
    - Telecom
    - Telecom VIC
    - Sandstone Quarry
    - Water Utility 1
    - Water Utility 2
    - Water Utility 6 VIC

3.6. Baseline logic

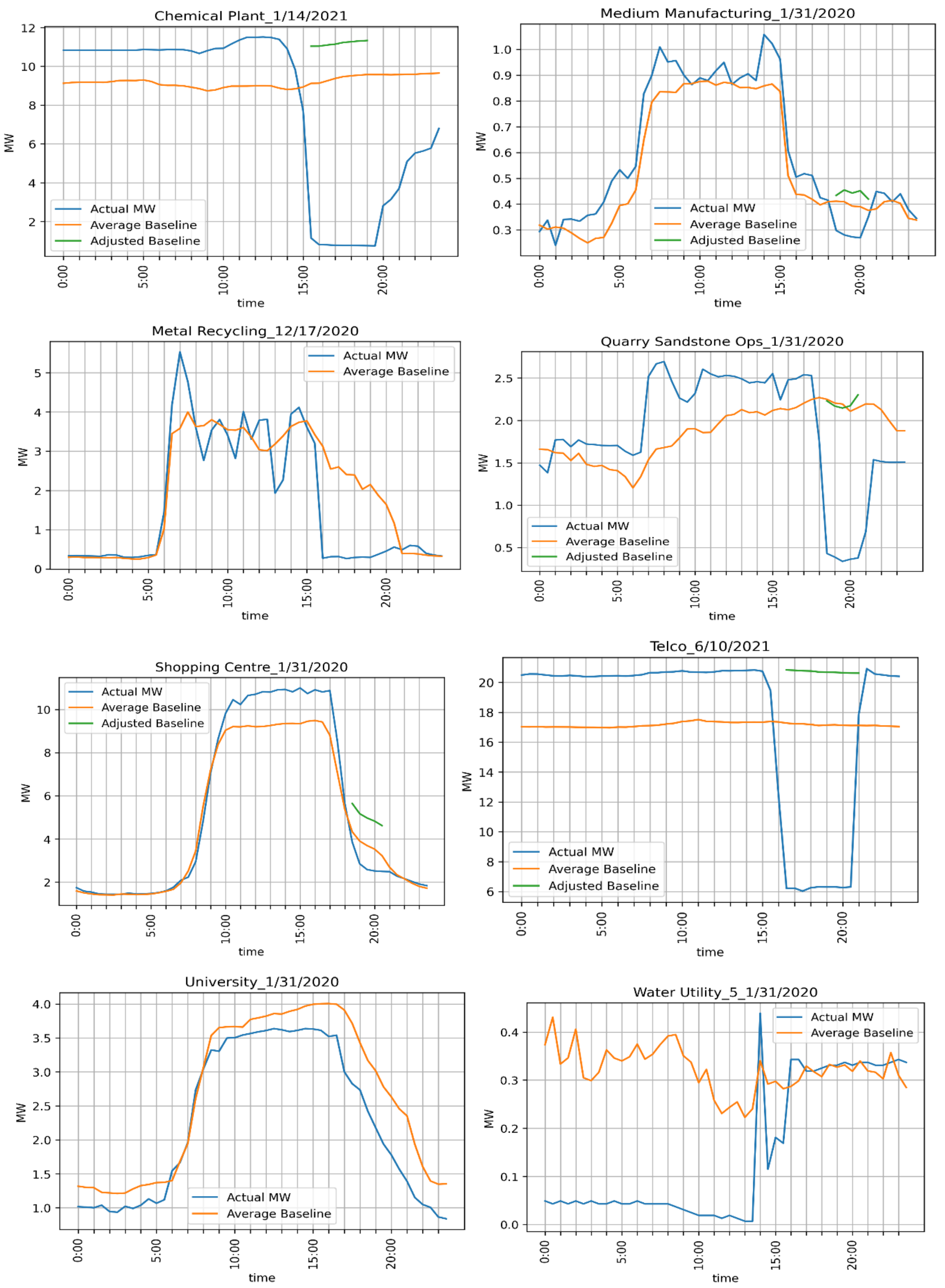
The baseline logic currently in use is one of the averaging methods. Averaging methods are often called representative day methods or High X of Y methods in that:

"A High X of Y baseline considers the Y most recent days preceding an event and uses the data from the X days with the highest load within those Y days to calculate the baseline [9]."

According to the data provided, the baseline is formed as follows:

- Calculation of average of the last ten business days (*public holidays are not counted for the average)
- No ranking of days.
- For every interval, the average is calculated based on those ten business days.
- An adjustment factor is determined by the difference between the metered demand level and the baseline average in a three-hour period starting four hours before the event starting time.
- The adjustment factor creates an adjusted baseline with a 20% increase cap but no decrease cap.

This means Y=10 and X=10. From now on, we refer to this as **'Average Baseline'**. The **'Adjusted Baseline'** results are shown in Figure 4. A baseline adjustment (sometimes called a "day of adjustment") is made based on data from the day of the event. Most baseline adjustments use a timeframe of 2 to 4 hours before the event. More than one hour is needed to represent the difference, and four or more hours may consider conditions too far from the event to be representative. The actual load over this time period is compared to the load estimated by the baseline over the same time period and is used to calculate the appropriate adjustment [13]. In Figure 4, the adjustment timeframe is 1 to 4 hours prior to the event, and its magnitude is limited by a 20% increase. As can be seen from the graphs, some portfolios satisfy the 20% increase cap condition.


Chemical Plant_1/14/2021

Medium Manufacturing_1/31/2020

Metal Recycling_12/17/2020

Quarry Sandstone Ops_1/31/2020

Shopping Centre_1/31/2020

Telco_6/10/2021

University_1/31/2020

Water Utility_5_1/31/2020

Figure 4. A comparison between the actual power consumption (MW), average baseline and adjusted baseline formula for eight portfolios[1].

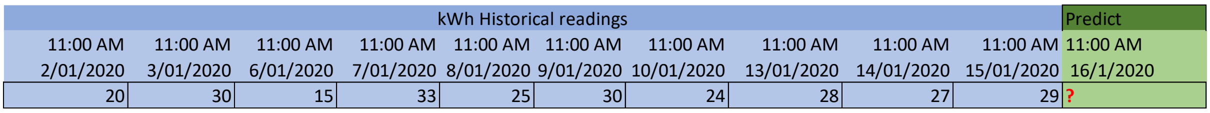
[1] In some sub-figures adjusted baseline is not shown, since there is not enough temporal data to calculate the adjusted baseline. To calculate the adjusted baseline, it requires more than five hours of data before the event.

4. Application Case Study: The AGL Dataset

This section will introduce different ML techniques for the AGL dataset and discuss their prediction results and the demand response performance compared to the average baseline model. The dataset contains daily power consumption in kWh for 12 C&I customers. It includes the power consumption of an event day (actual DR event), which has been excluded during the training and testing of the prediction models presented.

4.1. Predictive models and evaluation methods

In this section, we apply different ML techniques for DR prediction and compare them with the average baseline model (averaging method). For all predictive models, to predict the power consumption at a particular time stamp, e.g., 11:00, the model uses ten historical data points, as illustrated in Figure 5.



kWh Historical readings										Predict
11:00 AM	11:00 AM	11:00 AM	11:00 AM	11:00 AM	11:00 AM	11:00 AM	11:00 AM	11:00 AM	11:00 AM	11:00 AM
2/01/2020	3/01/2020	6/01/2020	7/01/2020	8/01/2020	9/01/2020	10/01/2020	13/01/2020	14/01/2020	15/01/2020	16/1/2020
20	30	15	33	25	30	24	28	27	29	?

Figure 5. Example of using historical data (10/10) in the prediction and excluding weekend days for DR calculation.

In our experiments, besides the average baseline model, the following predictive models are used, polynomial linear regression (a nonlinear predictor, denoted as NLR), support vector regression, denoted as (SVR), and Bayesian linear regression, denoted as (BLR). The technical details of these models are presented in the Appendix. For evaluation of all models on the AGL dataset, the following accuracy metric is used:

$$Accuarcy\ (ACC) = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(|\hat{y}_i - y_i|)}{\frac{1}{n}\sum_{i}^{n} y_i} \qquad (3)$$

where $\hat{y}_i$ is the predicted power consumption, y_i is the real power consumption, and n is the number of prediction samples that depends on the time series length in the test set.

For each model, the following cases are explored to verify the effectiveness of weather-related data on the demand prediction:

- Case 0: energy consumption,
- Case 1: energy consumption and temperature.

4.2. Results

This section presents the comparison results of the average baseline and prediction models in terms of prediction accuracy (ACC) for the AGL dataset. The prediction accuracy for each ML model compared to average baseline and for both cases are discussed.

4.2.1. Case 0 results

The section presents the accuracy results of predictive and average baseline models using only electricity data, not temperature. For all following models, the historical electricity data over the same time stamp is applied to predict the future event at the same time stamp.

Support vector regression results (SVR)

Figure 6 shows the percentage accuracy results on the test set for eleven (11) C&I customers compared to the average (shown as ‘Avg. baseline’ in plots) baseline model. For some C&I customers, such as (Chemical plant, Telecom, Telecom VIC, Water Utility 2, and Water Utility 5), the SVR performs better than the average baseline model. The overall accuracy on eleven (11) C&I customers shows that the average baseline model has better performance with an accuracy of 74%, while SVR has an accuracy of 59%.

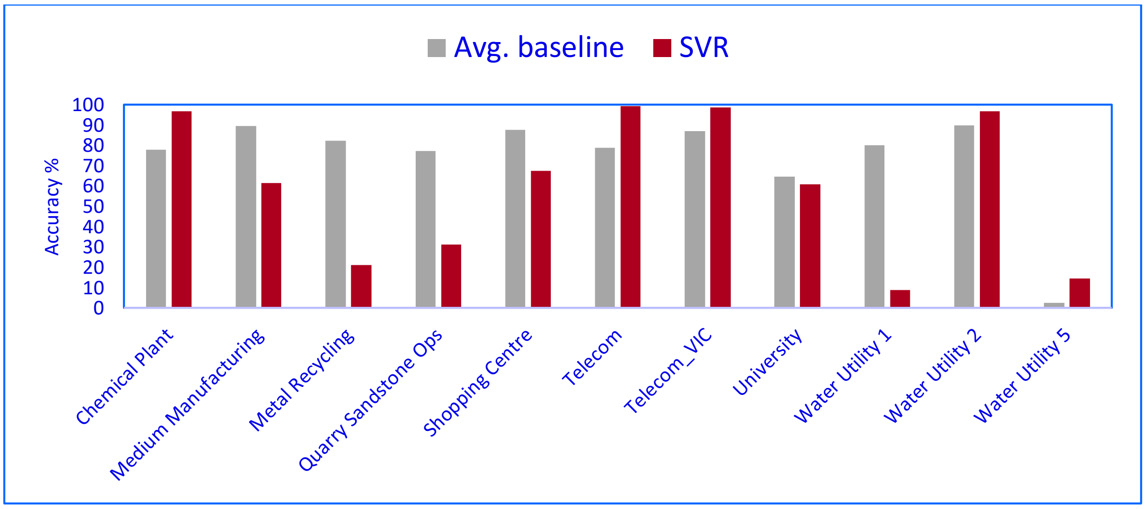


Figure 6. Testing accuracy performance for average baseline and SVR model for case 0.

Bayesian linear regression results (BLR)

Figure 7 presents the accuracy results of the BLR model compared to the average baseline model. The BLR model shows better accuracy for 6 out of 11 C&I customers compared to the average baseline. The overall accuracy of the BLR model is 78%, with a 4% margin compared to the average baseline model.

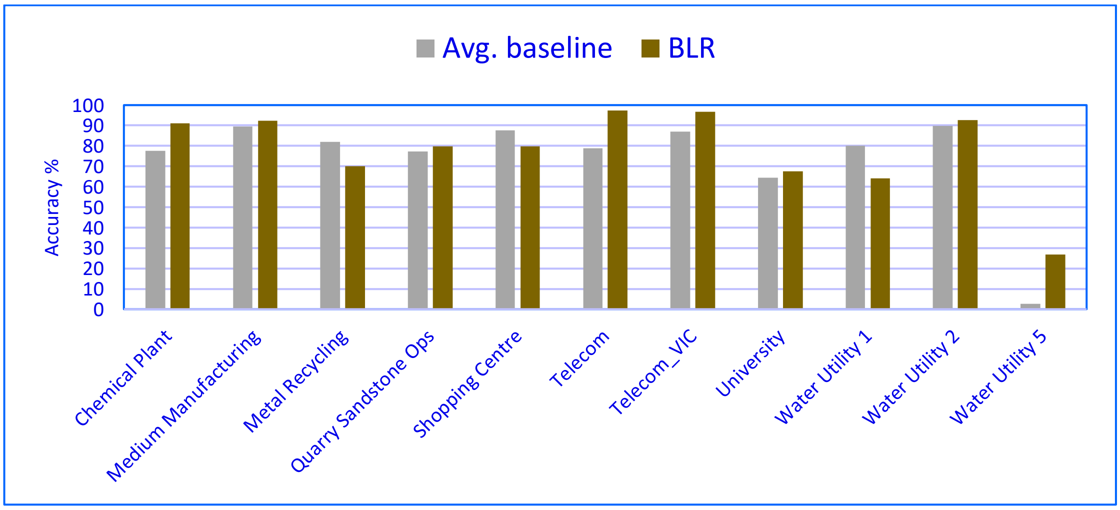


Figure 7. Testing accuracy performance for average baseline and BLR model for case 0.

Polynomial regression results (NLR)

Figure 8 reports the accuracy of the NLR model compared to the baseline model. The NLR achieves better accuracy for 8 out of 11 C&I customers compared to the average baseline model. There is a 5% margin increase in the overall accuracy (97%) compared to the average baseline model (74%).

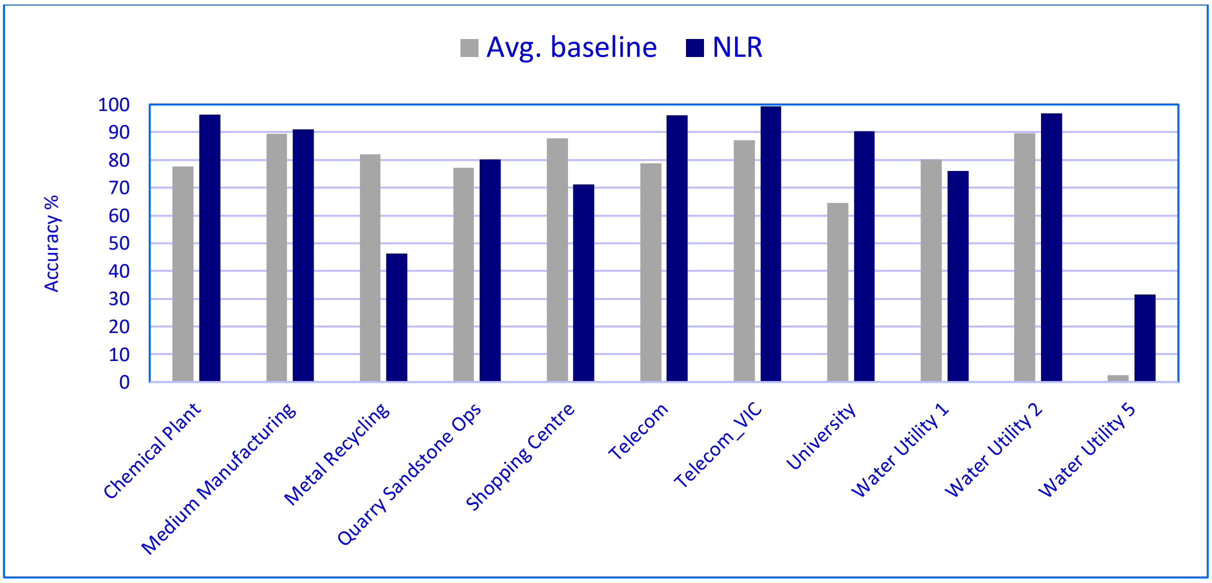


Figure 8. Testing accuracy performance over baseline and NLR model for case 0.

In general, we can observe that for the following C&I customers, ML performs better than the average baseline model. Thus, we highly recommend using ML for baseline calculation (e.g. NLR) for the following C&I customers:

- Chemical Plant
- Telecom
- Telecom VIC
- Water utility 2
- Water utility 5

4.2.2. Case 1 results

This section presents the accuracy results of predictive and baseline models using electricity data and maximum temperature data as an exogenous variables[2]. For all the following models, the historical electricity data with maximum daily temperature over the same time stamp is applied to predict the future event at the same time stamp. Figure 9 shows the results of the ML models using daily maximum temperature as the exogenous variable for the predictive model. In general, we see there is not much improvement in model performance compared to case 0. The only difference in the performance is observed for the Shopping Centre. Both the NLR and SVR accuracies have improved by including the maximum temperature, as shown in Figure 10. The general accuracy of the AGL dataset for case 0 and case 1 is shown in Table 5.

[2] Exogenous variable is a variable that is not affected by other variables in the model.

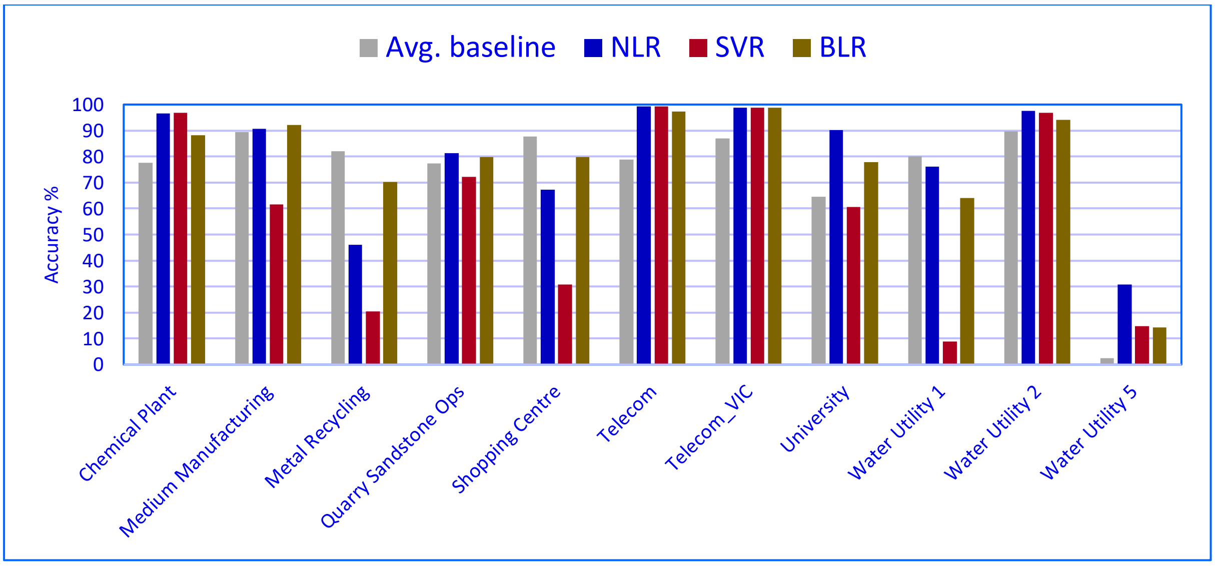


Figure 9. Testing accuracy performance over baseline and NLR, SVR, and BLR models for case 1.

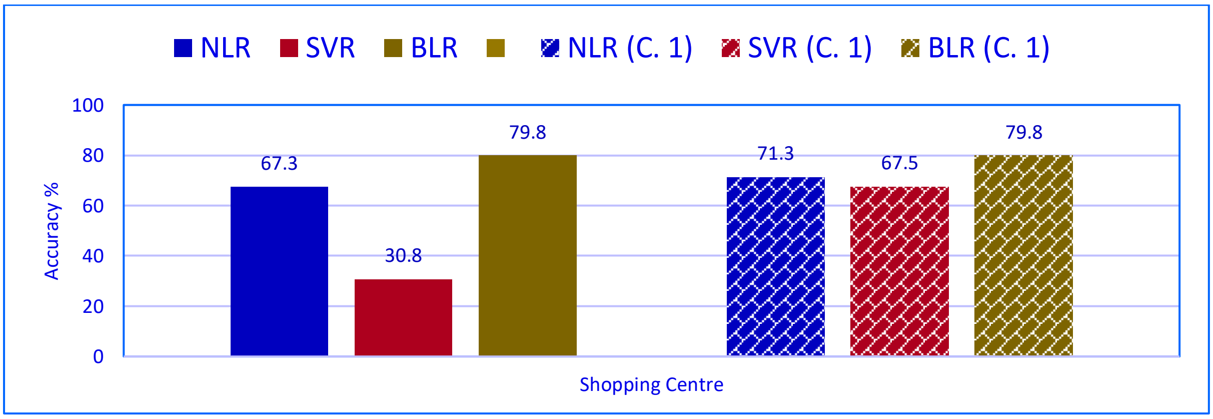


Figure 10. The accuracy of the shopping centre for case 0 and case 1 (denoted by C. 1).

Table 5. AGL overall accuracy of ML models and baseline for both case 0 and case 1.

Model	Overall Accuracy %	
	Case 0	***Case 1***
Avg. Baseline	74	74.0
NLR	79	79.4
SVR	59	60.0
BLR	78	77.8

4.3. Demand response analysis

This section presents an analysis of the ML models for the DR event compared to average baseline models. We choose the best model, i.e., NLR, based on the accuracy presented in the previous section for DR analysis and compared it with the average baseline model. Figure 11 shows the DR prediction of the NLR model and baseline model compared to the actual event consumption.

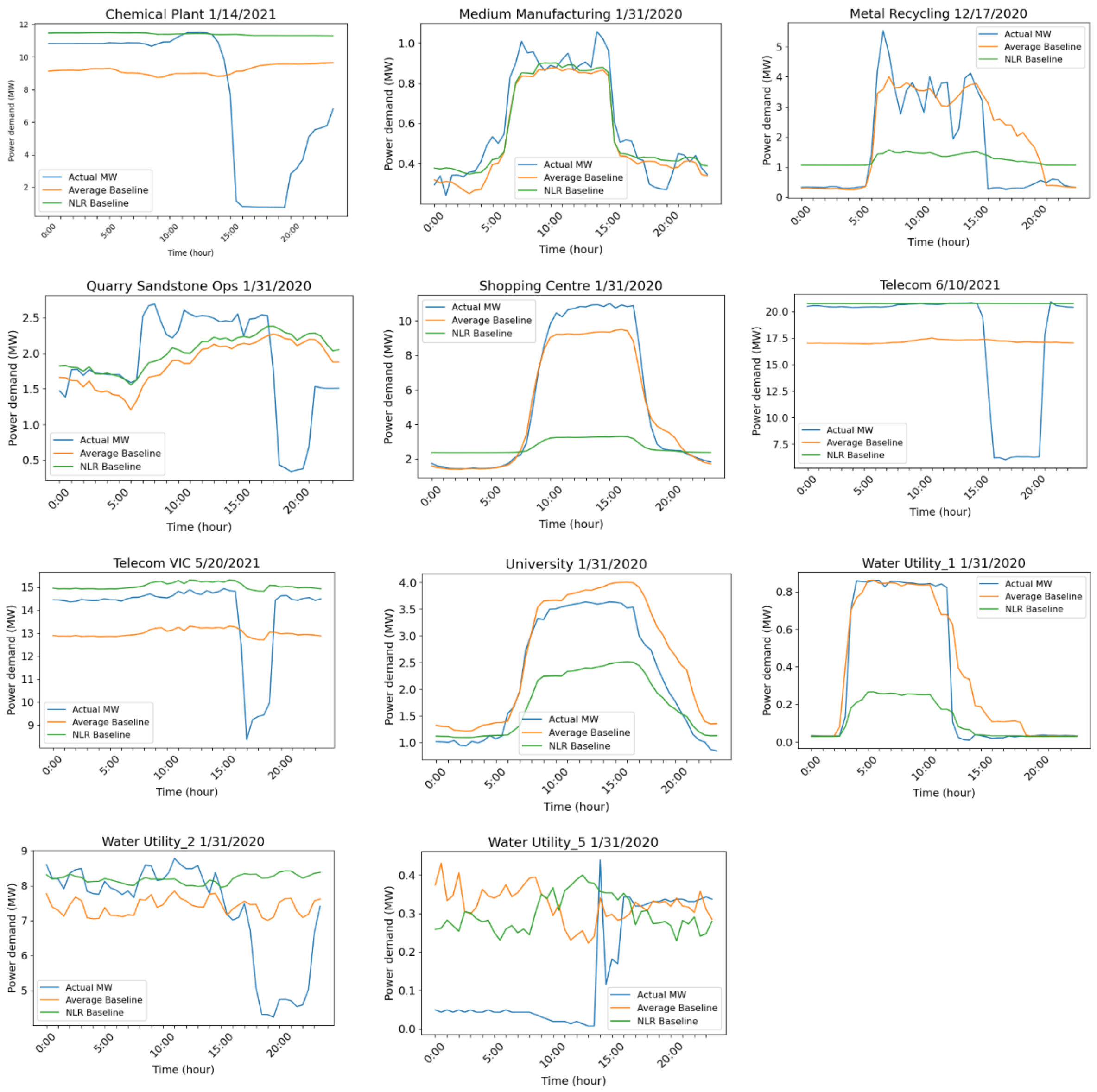
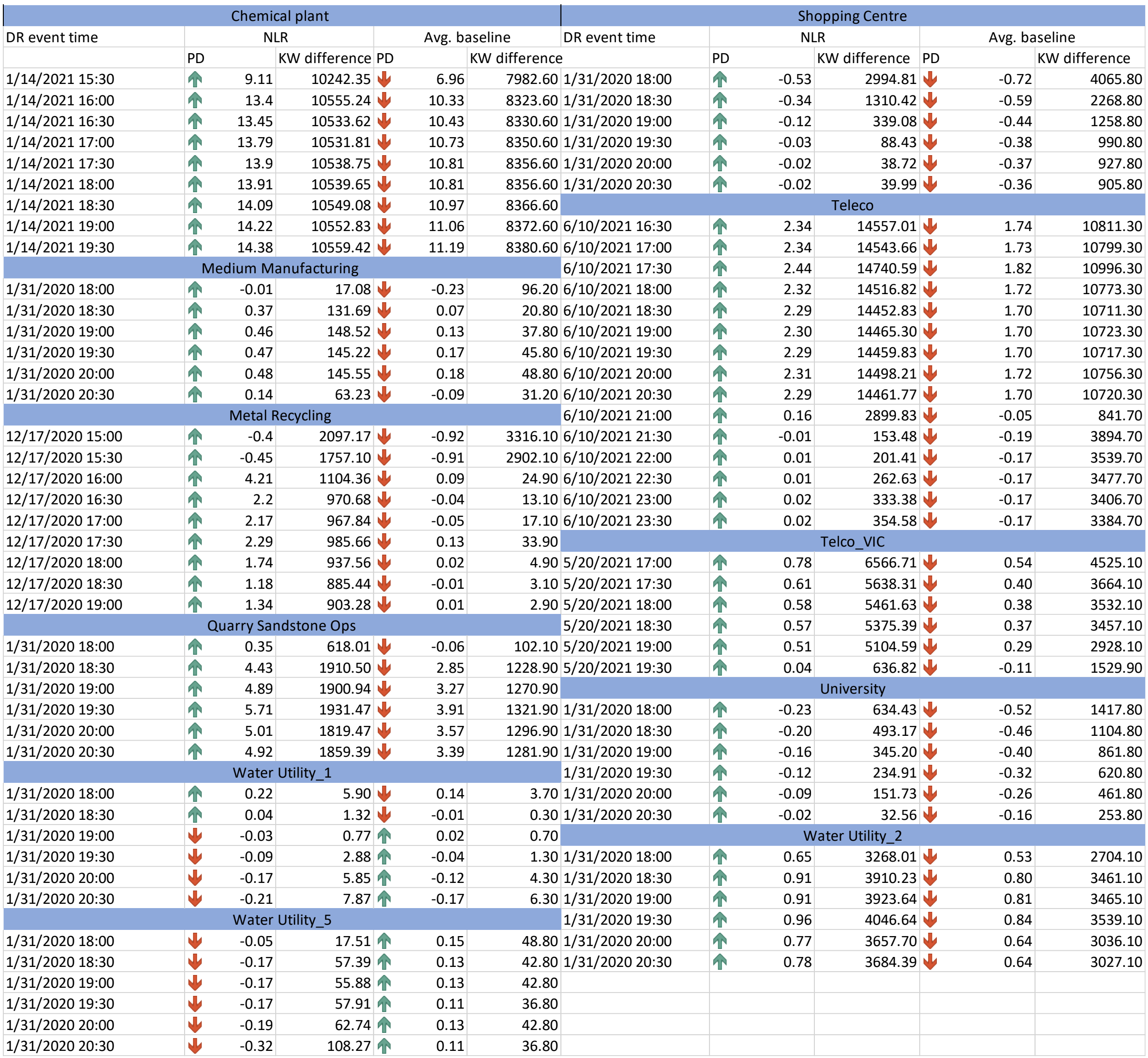


Figure 11. Demand response predictions for NLR and baseline on AGL dataset with 11 C&I customers.

We report the demand response percentage difference as follows:

$$DR\ Percentage\ Differnce\ (PD) = \frac{Predicated\ value - Actual\ DR\ value}{Actual\ DR\ value} \qquad (4)$$

If the predicted NLR value or the baseline is lower than the actual event value, the demand variance will be negative. Otherwise, it is positive. The positive and negative values reflect if the C&I customer will receive positive or negative benefits by using an ML baseline calculation. Figure 12 shows the difference in DR for the Chemical Plant, Medium Manufacturing, and Metal Recycling profiles for NLR and the average baseline model. The symbol ↑ shows that NLR has a more positive benefit than the average baseline model.



Chemical plant						
DR event time	NLR			Avg. baseline		
	PD		KW difference	PD		KW difference
1/14/2021 15:30	↑	9.11	10242.35	↓	6.96	7982.60
1/14/2021 16:00	↑	13.4	10555.24	↓	10.33	8323.60
1/14/2021 16:30	↑	13.45	10533.62	↓	10.43	8330.60
1/14/2021 17:00	↑	13.79	10531.81	↓	10.73	8350.60
1/14/2021 17:30	↑	13.9	10538.75	↓	10.81	8356.60
1/14/2021 18:00	↑	13.91	10539.65	↓	10.81	8356.60
1/14/2021 18:30	↑	14.09	10549.08	↓	10.97	8366.60
1/14/2021 19:00	↑	14.22	10552.83	↓	11.06	8372.60
1/14/2021 19:30	↑	14.38	10559.42	↓	11.19	8380.60
Medium Manufacturing						
1/31/2020 18:00	↑	-0.01	17.08	↓	-0.23	96.20
1/31/2020 18:30	↑	0.37	131.69	↓	0.07	20.80
1/31/2020 19:00	↑	0.46	148.52	↓	0.13	37.80
1/31/2020 19:30	↑	0.47	145.22	↓	0.17	45.80
1/31/2020 20:00	↑	0.48	145.55	↓	0.18	48.80
1/31/2020 20:30	↑	0.14	63.23	↓	-0.09	31.20
Metal Recycling						
12/17/2020 15:00	↑	-0.4	2097.17	↓	-0.92	3316.10
12/17/2020 15:30	↑	-0.45	1757.10	↓	-0.91	2902.10
12/17/2020 16:00	↑	4.21	1104.36	↓	0.09	24.90
12/17/2020 16:30	↑	2.2	970.68	↓	-0.04	13.10
12/17/2020 17:00	↑	2.17	967.84	↓	-0.05	17.10
12/17/2020 17:30	↑	2.29	985.66	↓	0.13	33.90
12/17/2020 18:00	↑	1.74	937.56	↓	0.02	4.90
12/17/2020 18:30	↑	1.18	885.44	↓	-0.01	3.10
12/17/2020 19:00	↑	1.34	903.28	↓	0.01	2.90
Quarry Sandstone Ops						
1/31/2020 18:00	↑	0.35	618.01	↓	-0.06	102.10
1/31/2020 18:30	↑	4.43	1910.50	↓	2.85	1228.90
1/31/2020 19:00	↑	4.89	1900.94	↓	3.27	1270.90
1/31/2020 19:30	↑	5.71	1931.47	↓	3.91	1321.90
1/31/2020 20:00	↑	5.01	1819.47	↓	3.57	1296.90
1/31/2020 20:30	↑	4.92	1859.39	↓	3.39	1281.90
Water Utility_1						
1/31/2020 18:00	↑	0.22	5.90	↓	0.14	3.70
1/31/2020 18:30	↑	0.04	1.32	↓	-0.01	0.30
1/31/2020 19:00	↓	-0.03	0.77	↑	0.02	0.70
1/31/2020 19:30	↓	-0.09	2.88	↑	-0.04	1.30
1/31/2020 20:00	↓	-0.17	5.85	↑	-0.12	4.30
1/31/2020 20:30	↓	-0.21	7.87	↑	-0.17	6.30
Water Utility_5						
1/31/2020 18:00	↓	-0.05	17.51	↑	0.15	48.80
1/31/2020 18:30	↓	-0.17	57.39	↑	0.13	42.80
1/31/2020 19:00	↓	-0.17	55.88	↑	0.13	42.80
1/31/2020 19:30	↓	-0.17	57.91	↑	0.11	36.80
1/31/2020 20:00	↓	-0.19	62.74	↑	0.13	42.80
1/31/2020 20:30	↓	-0.32	108.27	↑	0.11	36.80

Shopping Centre						
DR event time	NLR			Avg. baseline		
	PD		KW difference	PD		KW difference
1/31/2020 18:00	↑	-0.53	2994.81	↓	-0.72	4065.80
1/31/2020 18:30	↑	-0.34	1310.42	↓	-0.59	2268.80
1/31/2020 19:00	↑	-0.12	339.08	↓	-0.44	1258.80
1/31/2020 19:30	↑	-0.03	88.43	↓	-0.38	990.80
1/31/2020 20:00	↑	-0.02	38.72	↓	-0.37	927.80
1/31/2020 20:30	↑	-0.02	39.99	↓	-0.36	905.80
Teleco						
6/10/2021 16:30	↑	2.34	14557.01	↓	1.74	10811.30
6/10/2021 17:00	↑	2.34	14543.66	↓	1.73	10799.30
6/10/2021 17:30	↑	2.44	14740.59	↓	1.82	10996.30
6/10/2021 18:00	↑	2.32	14516.82	↓	1.72	10773.30
6/10/2021 18:30	↑	2.29	14452.83	↓	1.70	10711.30
6/10/2021 19:00	↑	2.30	14465.30	↓	1.70	10723.30
6/10/2021 19:30	↑	2.29	14459.83	↓	1.70	10717.30
6/10/2021 20:00	↑	2.31	14498.21	↓	1.72	10756.30
6/10/2021 20:30	↑	2.29	14461.77	↓	1.70	10720.30
6/10/2021 21:00	↑	0.16	2899.83	↓	-0.05	841.70
6/10/2021 21:30	↑	-0.01	153.48	↓	-0.19	3894.70
6/10/2021 22:00	↑	0.01	201.41	↓	-0.17	3539.70
6/10/2021 22:30	↑	0.01	262.63	↓	-0.17	3477.70
6/10/2021 23:00	↑	0.02	333.38	↓	-0.17	3406.70
6/10/2021 23:30	↑	0.02	354.58	↓	-0.17	3384.70
Telco_VIC						
5/20/2021 17:00	↑	0.78	6566.71	↓	0.54	4525.10
5/20/2021 17:30	↑	0.61	5638.31	↓	0.40	3664.10
5/20/2021 18:00	↑	0.58	5461.63	↓	0.38	3532.10
5/20/2021 18:30	↑	0.57	5375.39	↓	0.37	3457.10
5/20/2021 19:00	↑	0.51	5104.59	↓	0.29	2928.10
5/20/2021 19:30	↑	0.04	636.82	↓	-0.11	1529.90
University						
1/31/2020 18:00	↑	-0.23	634.43	↓	-0.52	1417.80
1/31/2020 18:30	↑	-0.20	493.17	↓	-0.46	1104.80
1/31/2020 19:00	↑	-0.16	345.20	↓	-0.40	861.80
1/31/2020 19:30	↑	-0.12	234.91	↓	-0.32	620.80
1/31/2020 20:00	↑	-0.09	151.73	↓	-0.26	461.80
1/31/2020 20:30	↑	-0.02	32.56	↓	-0.16	253.80
Water Utility_2						
1/31/2020 18:00	↑	0.65	3268.01	↓	0.53	2704.10
1/31/2020 18:30	↑	0.91	3910.23	↓	0.80	3461.10
1/31/2020 19:00	↑	0.91	3923.64	↓	0.81	3465.10
1/31/2020 19:30	↑	0.96	4046.64	↓	0.84	3539.10
1/31/2020 20:00	↑	0.77	3657.70	↓	0.64	3036.10
1/31/2020 20:30	↑	0.78	3684.39	↓	0.64	3027.10

Figure 12. DR percentage difference (PD) and kW difference between the predicted value using NLR and the average baseline methods and the actual DR value, the green arrows indicate where there is a positive benefit.

5. Application Case Study: The Greater Western Water Dataset

This section shows the details of the Greater Western Water dataset, the predictive model, prediction results, and the demand response performance for the ML models used in comparison to the average baseline model.

5.1. Dataset description and evaluation method

The data was collected every 15 minutes from 28/05/2019 to 28/05/2021. Given the predictors, the inputs that are used to predict future demand will include ten historical demand data points. Before training, data cleaning is performed consumption days, days with undefined values, weekends, and days with power outages. We use the same evaluation accuracy (ACC) as presented in Section 4.1. The same predictive models in the previous section are used, which are polynomial linear regression (NLR), support vector regression (SVR), and Bayesian linear regression (BLR).

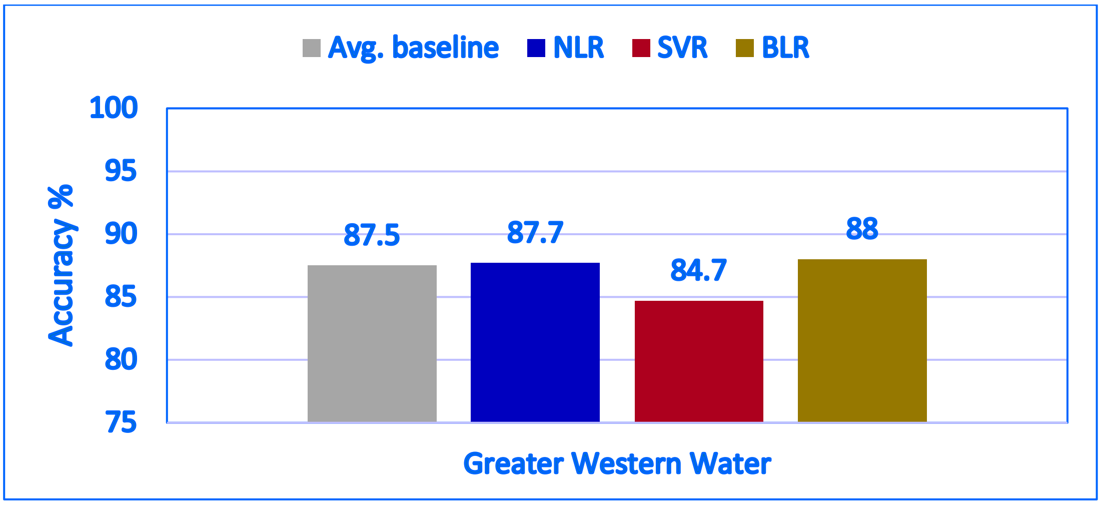
For each model, the following cases are explored to verify the effectiveness of weather-related data on the demand prediction:

- Case 0: energy consumption.
- Case 1: energy consumption and temperature.

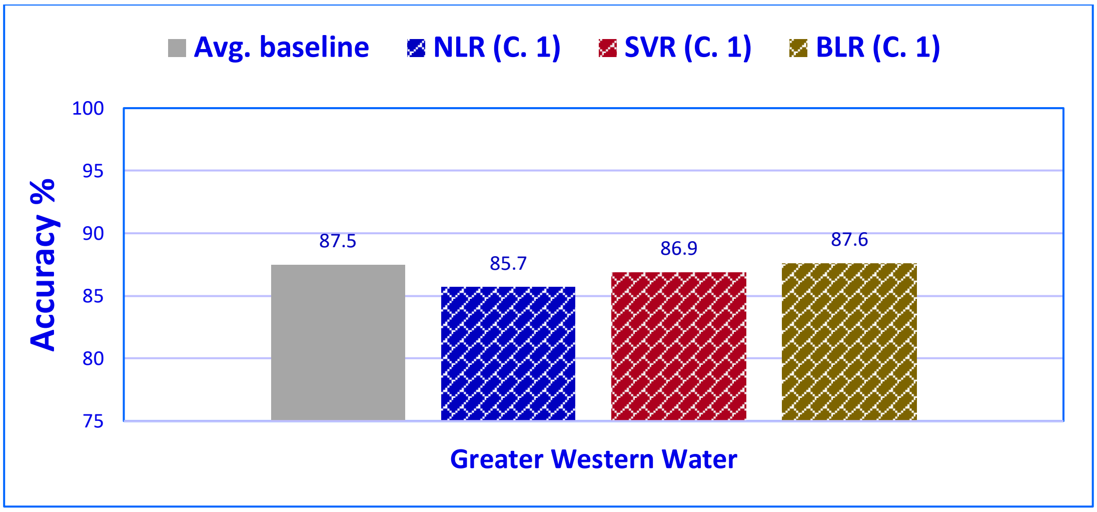
5.2. Results and DR analysis

In this section, the accuracy of the different predictive models and baseline models are reported. Figure 13 shows both cases' prediction accuracies of NLR, SVR, BLR, and the average baseline models. For case 0, as shown in Figure 13 (a), there is a slight difference in the performance between the ML predictive models and the average baseline model. However, the performance of the ML predictive models decreases when the maximum temperature is included, as shown in Figure 13 (b).

Figure 14 shows the DR predictions of the NLR, SVR, and BLRs models and the average baseline model on the data for 27/05/2021 compared to the actual event consumption.



(a). Case 0.



(b) Case 1 denoted by C. 1.

Figure 13. Test accuracy of the Greater Western Water dataset for different models for (a) case 0; and (b) case 1.

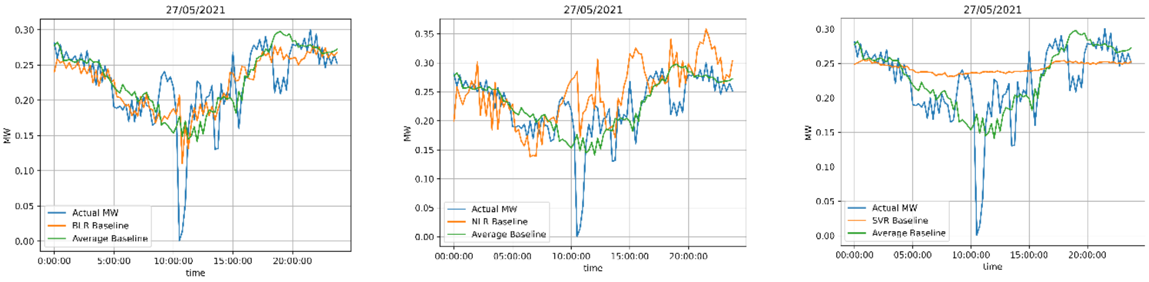


Figure 14. Demand response predictions for NLR and the average baseline on the Greater Western Water dataset.

6. Application Case Study: The Gippsland Water Dataset

This section presents details of the Gippsland data, the prediction model, the prediction results, and the demand response performance over the ML models used in comparison with to the average baseline model.

6.1. Dataset description and evaluation methods

The data was collected every 30 minutes from 01/04/2018 to 31/12/2020 and is shown in Figure 15. In addition to the demand data, the data includes exogenous variables, such as temperature, humidity, pressure, and precipitation, which are shown in Figure 16. Given the predictors, the inputs that are used to predict the future demand include historical demand data and some of the exogenous variables that are employed to help improve the prediction accuracy.

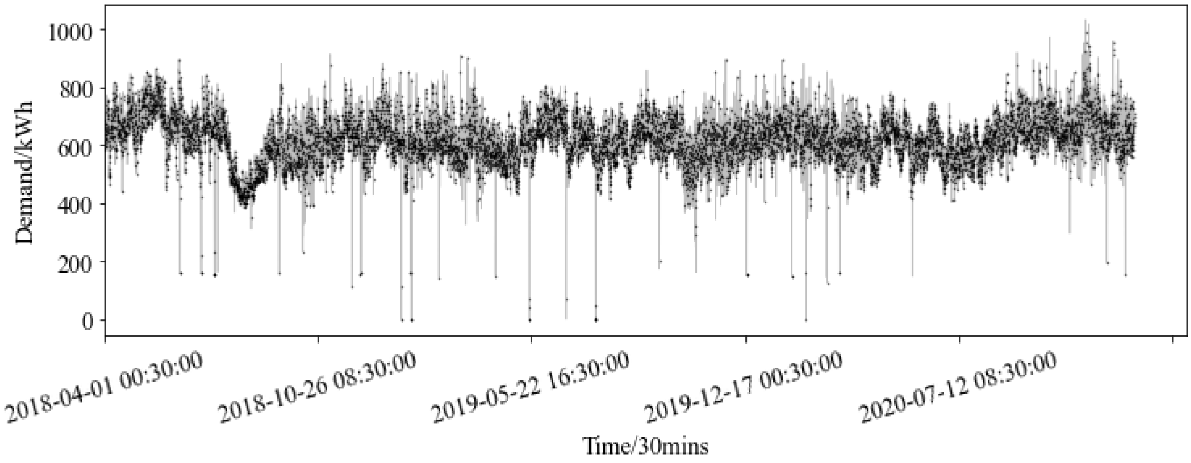


Figure 15: Demand data from 01/04/2018 to 31/12/2020.

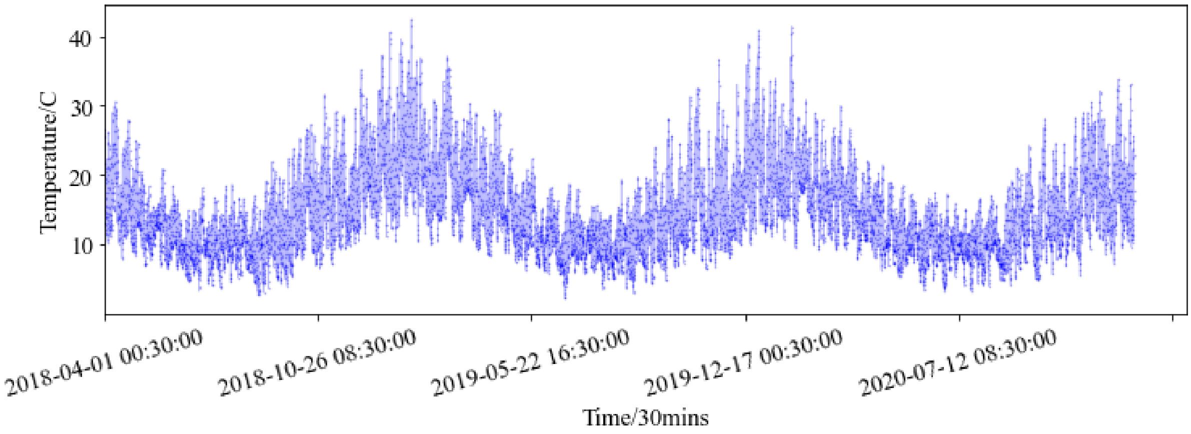


Figure 16: Temperature data from 01/04/2018 to 31/12/2020.

This case study used the data before 28/12/2020 (training set) to train the prediction model and the data from 28/12/2020 to 30/12/2020 (testing set) to test the trained model. To evaluate the accuracy of the prediction, this study chose two typical evaluation metrics, i.e., mean absolute error (MAE) and root mean square error (RMSE). The mathematical formulations of these metrics are detailed in the Appendix. For both evaluation methods, a smaller value represents better performance.

- **MAE**: measures the average magnitude of the errors in a set of predictions. It is the average over the samples of the absolute differences between predictions and actual observations.
- **RMSE**: a quadratic scoring rule that also measures the average magnitude of the error. It is the square root of the average of squared differences between predictions and actual observations.

6.2. Predictive modelling

Instead of focusing on the most accurate prediction, this study aims to use the predictors that can have better prediction performance in comparison to the baseline models for DR analysis. The linear regression (LR) and polynomial regression (a nonlinear predictor, denoted as NLR) are used as the predictors. The details of these two predictors are presented in the Appendix.

For each of the predictors used here, we designed two models for the demand prediction:

- **Model 1**: like the average baseline, it uses the same historical time stamps to predict the same future time, e.g., using the 10:00 historical data to predict 10:00 am in the future with a time window (the number of historical days used as inputs).

- **Model 2**: uses the time stamps before the predicted time stamps as inputs with a time window (the number of previous time stamps used as inputs), e.g., using 7:00, 7:30, 8:00, 8:30, 9:00, and 9:30 data to predict 10:00 for the same day.

For each model, the following cases are explored to verify the effectiveness of the weather-related data on the demand prediction:

- Case 0: energy consumption.

- Case 1: energy consumption and temperature.

- Case 2: energy consumption, temperature, daily precipitation rate and solar radiation information.

6.3. Results

In this section, the results are discussed including the prediction error (MAE and RMSE) for Models 1 and 2. They are also compared with the average baseline model. The time window (TW) is set as 4 (TW=4), 7 (TW=7), 10 (TW=10), and 14 (TW=14) for all models. The time window (TW) represents the days in Model 1 and the time stamps before the predicted value in Model 2. In the following results, M1 and M2 denote Model 1 and Model 2, respectively.

6.3.1. Prediction result over Model 1

For Model 1, the historical electricity consumption data over the same time stamp is applied to predict the future event at the same time stamp. We compare the MAE and RMSE of Case 0, Case 1 and Case 2 in Model 1 with the average baseline over LR and NLR. Figure 17 and Figure 18 report the results of MAE and RMSE over average baseline, LR and NLR in M1 across Case 0, Case 1, and Case 2 for training and testing sets. For the average baseline model, TW=14 (days) leads to the best performance on both training and testing sets. For LR, Case0 with TW=10 has the lowest MAE and RMSE (i.e., the highest accuracy) over training and testing sets. For NLR, Case 2 with TW=7 leads to the best performance for both MAE and RMSE on training and testing sets.

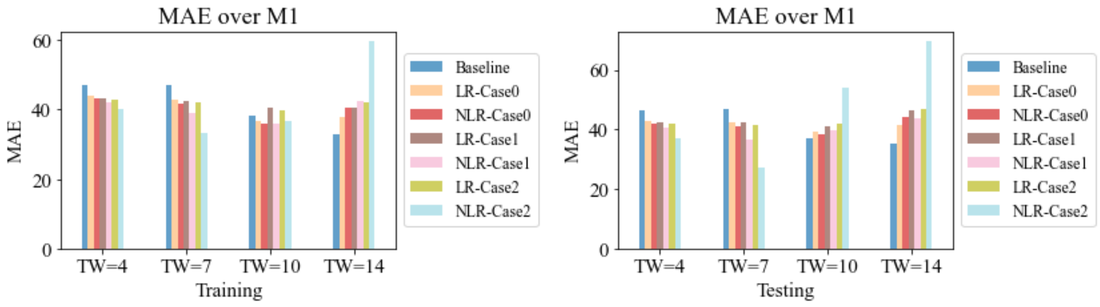


Figure 17: The training and testing MAE performance for average baseline and Model 1.

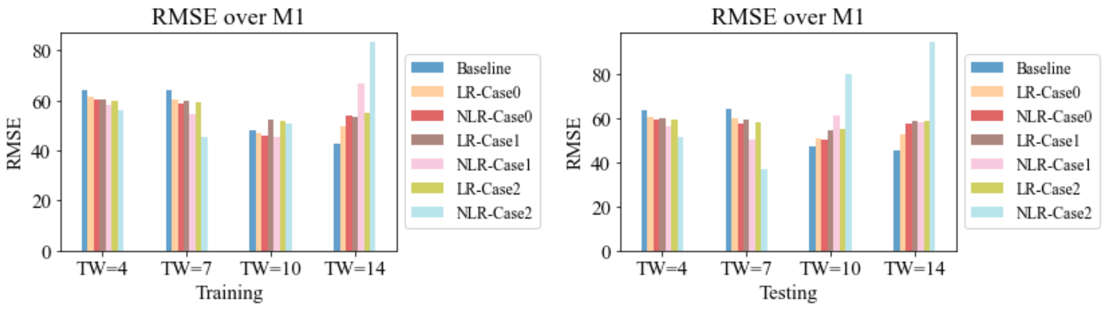


Figure 18: The training and testing RMSE performance for average baseline and Model 1.

By comparing the best performance over the average baseline, LR-Case 0 with TW=10 and NLR-Case 2 with TW=7:

- The average baseline model and NLR-Case2 with TW=7 led to higher accuracy than LR-Case 0 with TW=10 on training and testing sets for MAE and RMSE.
- NLR-Case 2 with TW=7 only has lower MAE and RMSE than the average baseline model on the testing set, while on the training set, it is not comparative to the average baseline model.

Neither LR nor NLR is comparative to the average baseline model, which indicates that using the energy data at the same time stamps without considering the continuity and dependency of the time series does not work well for this prediction problem.

6.3.2. Prediction result over Model 2

Figure 19 and Figure 20 present the MAE and RMSE results over the average baseline model, LR and NLR in Model 2 over different cases. Here the results of the average baseline model are the same as in Model 1, i.e., using the same time stamps of the previous days' data to obtain the predicted event. The results show that:

- The average baseline model has worse performance (i.e., higher MAE and RMSE) than LR and NLR over any case and any time window setting for training and testing sets.
- With the previous time stamps as the inputs, LR and NLR lead to higher accuracy on both training and testing sets, even only with electricity data, i.e., Case 0.
- By adding more factors like Case 0, Case 1, and Case 2, the MAE and RMSE values on the training and testing sets are further improved for both LR and NLR over each TW setting, except NLR-Case 2 on MAE and RMSE and LR-Case 2 on MAE when TW=10 and TW=14.

- By increasing the time window TW, the MAE and RMSE on training and testing for LR and NLR start to increase, i.e., the accuracy decreases. TW=10 leads to the best prediction performance for both LR and NLR.
- NLR-Case1 with TW=10 (i.e., the best performance of NLR across all cases and TW settings) outperforms LR-Case1 with TW=10 (i.e., the best LR performance across all cases and TW settings) over training and testing sets for both MAE and RMSE.

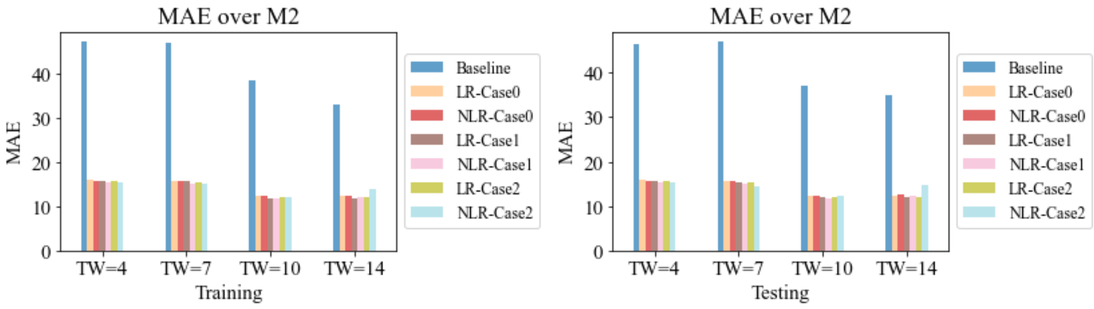


Figure 19: The training and testing MAE performance over average baseline and Model 2.

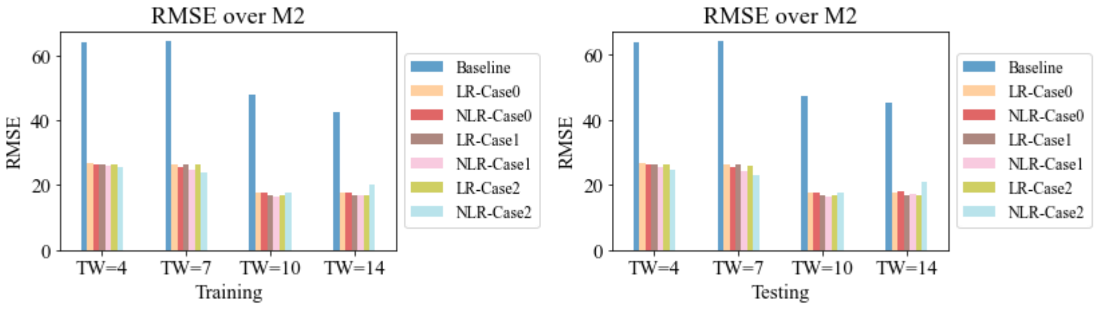


Figure 20: The training and testing RMSE performance over average baseline and Model 2.

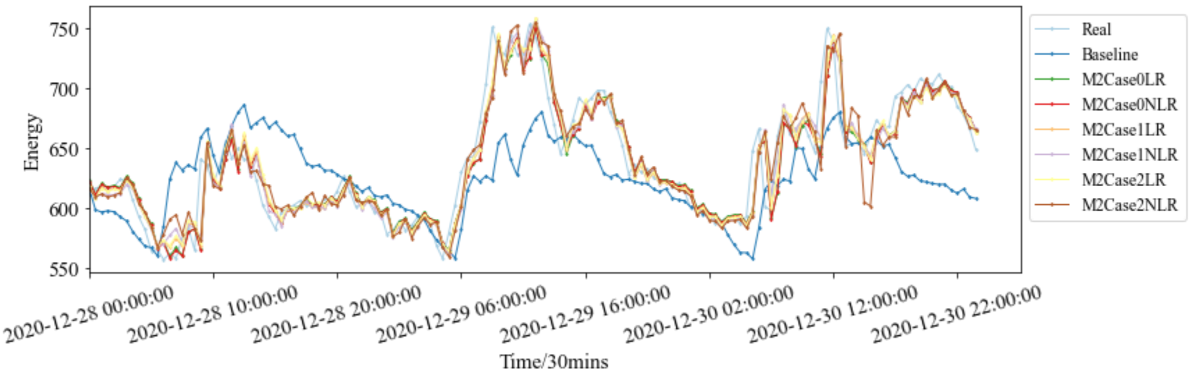


Figure 21: The comparison of the actual and predicted values over average baseline, LR, and NLR over cases in Model 2.

Since LR and NLR lead to better performance than the average baseline model over all the TW settings and cases, we visualise the predicted values on the testing set to see how they differ from the actual values, as presented in Figure 21. This visualisation tells that the predicted values of the Baseline model are more significantly different from the actual values of LR and NLR.

6.3.3. Demand response analysis

To analyse the demand response, we choose the best results from Model 2 for both LR and NLR, i.e., LR-Case1 and NLR-Case1 with TW=10 and compare them with the Baseline model. We use the demand difference, i.e., the difference between the predicted value or the baseline and the actual value, to analyse the demand response over the prediction horizon. If the predicted value or the baseline is lower than the real value, the demand difference will be negative. Otherwise, it is positive.

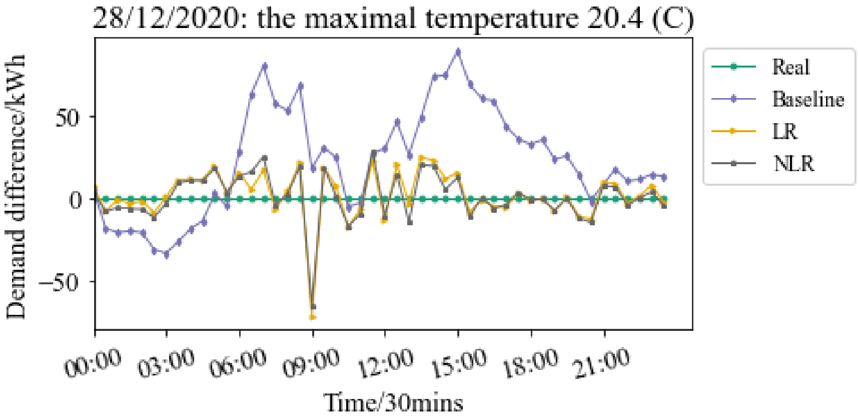


Figure 22: Demand difference (the prediction values of average baseline/LR/ NLR - the real values) over 28/12/2020.

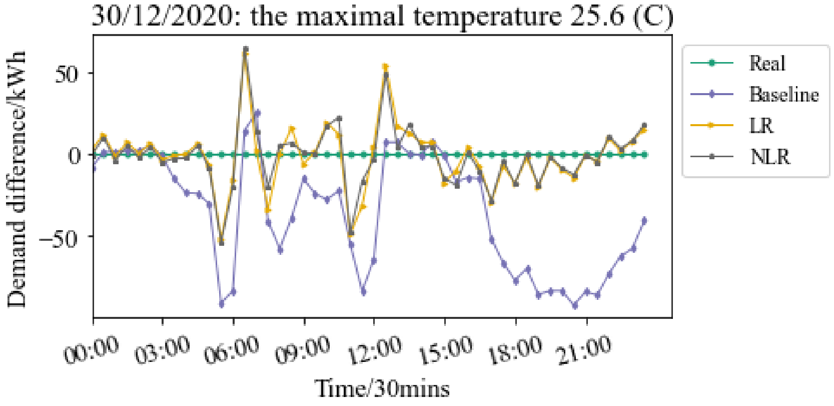


Figure 23: Demand difference (the prediction values of average baseline/LR/ NLR - the real values) over 29/12/2020.

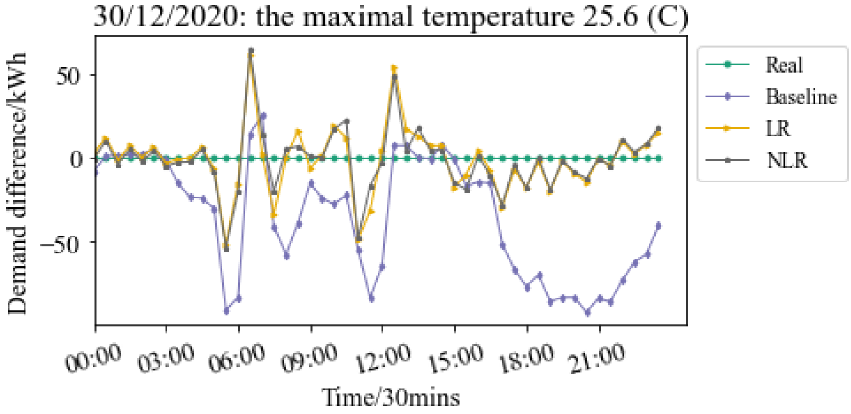


Figure 24: Demand difference (the prediction values of average baseline/LR/ NLR - the real values) over 30/12/2020.

Figure 22, Figure 23 and Figure 24 show the demand difference over the average baseline, LR, and NLR for 28/12/2020, 29/12/2020, and 30/12/2020. The green line represents the demand difference of the real values (all zeros). The results indicate that in most predicted values, LR and NLR have smaller demand differences than the average baseline model, which means their prediction is more accurate. Figure 22 illustrates the demand difference obtained by the real/LR/NLR values and the average baseline values. The purple line represents the demand difference of the average baseline (all zeros). It shows that the demand differences over the real values, LR and NLR have similar changes and are close to each other, which further verifies that the prediction of LR and NLR are quite close to the real values.

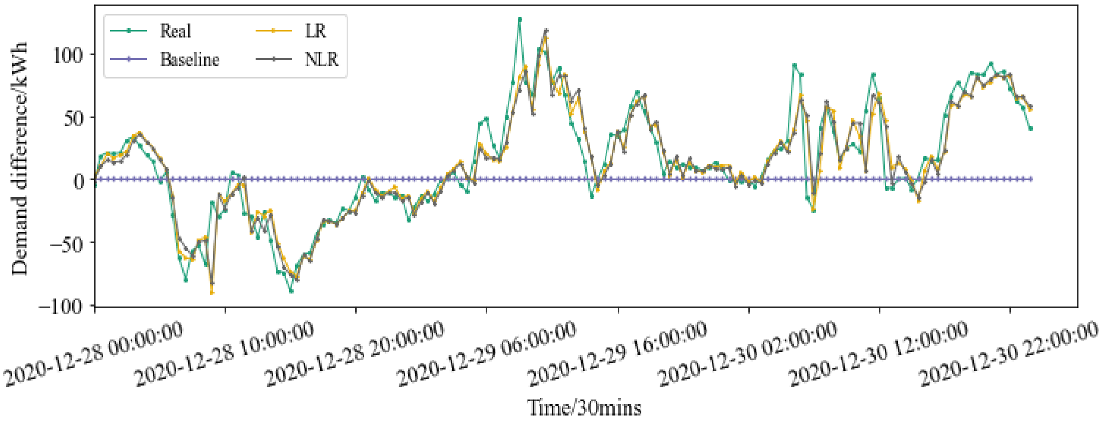


Figure 25: Demand difference (the real/LR/ NLR values - the average baseline values) over the testing days.

Table 6: The demand difference (i.e., the prediction values of average baseline/LR/NLR – the real values) percentage (i.e., demand differences divided by the real values).

Time	28/12/2020			29/12/2020			30/12/2020		
	Baseline	LR	NLR	Baseline	LR	NLR	Baseline	LR	NLR
0:00	0.81%	1.12%	0.51%	5.64%	3.40%	3.10%	-1.32%	0.44%	-0.09%
0:30	-2.96%	-1.34%	-1.35%	3.82%	-0.82%	-1.05%	0.17%	1.95%	1.66%
1:00	-3.42%	-0.14%	-0.94%	2.29%	-0.31%	-0.82%	0.24%	-0.22%	-0.63%
1:30	-3.24%	-0.54%	-1.05%	2.95%	1.27%	1.04%	0.32%	1.17%	0.81%
2:00	-3.40%	-0.32%	-1.12%	2.08%	-0.96%	-1.24%	0.33%	0.06%	-0.36%
2:30	-5.01%	-1.42%	-1.90%	0.16%	-1.02%	-1.54%	0.89%	1.09%	0.67%
3:00	-5.37%	0.10%	-0.53%	-0.16%	0.50%	0.27%	-0.26%	-0.57%	-0.86%
3:30	-4.35%	1.82%	1.50%	-0.99%	0.46%	0.17%	-2.56%	-0.17%	-0.48%
4:00	-3.17%	1.89%	1.79%	0.85%	3.28%	2.97%	-3.97%	-0.04%	-0.44%
4:30	-2.38%	1.91%	1.80%	1.70%	2.07%	2.11%	-4.09%	1.09%	0.82%
5:00	0.45%	3.50%	3.14%	-2.51%	-2.67%	-3.05%	-5.11%	-1.14%	-1.46%
5:30	-0.90%	0.45%	0.59%	-7.37%	-2.63%	-3.26%	-13.97%	-7.95%	-8.24%
6:00	5.07%	2.75%	2.36%	-7.62%	-4.44%	-5.01%	-12.47%	-2.36%	-3.02%
6:30	11.15%	0.91%	2.77%	-4.20%	-1.87%	-1.55%	2.36%	10.10%	10.73%
7:00	14.28%	3.00%	4.45%	-2.53%	-0.34%	-0.08%	4.22%	0.38%	2.33%
7:30	9.88%	-1.20%	-0.75%	-7.39%	-3.48%	-3.08%	-6.21%	-5.09%	-3.02%
8:00	9.06%	0.69%	0.42%	-10.97%	-3.37%	-3.38%	-8.45%	0.01%	0.76%
8:30	12.02%	3.77%	3.34%	-17.01%	-6.28%	-7.47%	-5.85%	2.40%	1.01%
9:00	2.83%	-11.22%	-10.11%	-10.57%	1.77%	1.11%	-2.26%	-0.94%	0.13%

9:30	4.76%	2.89%	2.90%	-9.20%	-1.46%	-1.99%	-3.59%	0.16%	0.04%
10:00	3.94%	1.11%	0.16%	-13.97%	-1.71%	-0.65%	-4.19%	2.86%	2.54%
10:30	-0.85%	-2.71%	-2.70%	-13.86%	1.60%	2.42%	-3.47%	1.84%	3.44%
11:00	-0.45%	-1.08%	-1.50%	-10.55%	0.27%	-1.31%	-7.77%	-6.90%	-6.79%
11:30	4.21%	3.43%	4.42%	-11.78%	-2.76%	-0.90%	-11.11%	-4.19%	-2.21%
12:00	4.59%	-1.99%	-1.63%	-9.01%	2.22%	2.03%	-8.66%	0.61%	-0.41%
12:30	7.17%	3.11%	2.24%	-6.15%	1.05%	2.51%	1.06%	8.03%	7.24%
13:00	4.10%	-0.48%	-2.25%	-4.54%	4.72%	**5.71%**	1.09%	2.55%	0.60%
13:30	7.82%	3.96%	3.23%	-2.13%	3.59%	3.91%	0.01%	1.93%	2.71%
14:00	12.16%	3.77%	3.21%	2.07%	4.96%	4.78%	-0.11%	1.08%	0.70%
14:30	12.57%	1.94%	0.83%	-0.02%	-1.35%	-0.67%	1.19%	1.08%	0.68%
15:00	15.25%	2.59%	2.12%	-1.77%	-0.80%	-1.39%	-0.14%	-2.66%	-2.28%
15:30	11.50%	-1.42%	-1.92%	-5.14%	-3.22%	-3.37%	-2.53%	-1.49%	-2.80%
16:00	10.03%	-0.17%	0.05%	-4.99%	0.52%	0.58%	-2.15%	0.63%	0.17%
16:30	9.77%	-0.86%	-1.06%	-5.73%	-2.09%	-2.57%	-2.28%	-1.14%	-1.59%
17:00	7.16%	-0.94%	-0.68%	-8.29%	-0.98%	-0.97%	-7.39%	-4.22%	-4.08%
17:30	5.92%	0.45%	0.57%	-9.92%	-1.07%	-1.33%	-9.51%	-1.04%	-0.59%
18:00	5.38%	-0.26%	-0.17%	-7.98%	1.68%	1.81%	-10.90%	-2.56%	-2.60%
18:30	5.84%	-0.12%	-0.11%	-6.14%	0.15%	-0.16%	-10.00%	-0.36%	-0.02%
19:00	3.82%	-1.36%	-1.26%	-4.47%	2.12%	2.63%	-12.06%	-2.80%	-2.70%
19:30	4.21%	0.10%	0.01%	-0.72%	2.84%	2.90%	-11.79%	-0.37%	-0.24%
20:00	2.30%	-1.82%	-2.03%	-2.24%	-1.77%	-1.53%	-11.81%	-1.35%	-1.22%
20:30	-0.31%	-2.05%	-2.37%	-1.45%	0.95%	1.39%	-12.90%	-2.03%	-1.75%
21:00	1.35%	1.50%	1.16%	-1.89%	-1.54%	-1.47%	-11.90%	-0.12%	-0.08%
21:30	2.86%	1.47%	1.11%	-1.52%	0.60%	1.12%	-12.23%	-0.66%	-0.70%
22:00	1.70%	-0.51%	-0.74%	-1.50%	-0.38%	-0.43%	-10.62%	1.36%	1.58%
22:30	1.86%	0.19%	0.02%	-1.10%	-0.27%	-0.03%	-9.15%	0.30%	0.50%
23:00	2.38%	1.31%	0.60%	-1.51%	0.28%	0.12%	-8.62%	1.18%	1.19%
23:30	2.16%	-0.45%	-0.77%	-2.18%	-0.62%	-0.80%	-6.26%	2.28%	2.72%
Mean	3.91%	0.36%	0.19%	-4.26%	-0.16%	-0.17%	-5.02%	-0.20%	-0.21%

Table 6 presents the demand difference percentage (demand difference divided by the real value) for the Baseline, LR, and NLR over the testing days for each time stamp. Since the tariff structure is not used to calculate the demand response benefit, we summarise the average value of each model across each day, from which an average value close to 0% is beneficial. NLR leads to smaller positive demand response influence on 28/12/2020, and LR has smaller negative influence on 29/12/2020 and 29/12/2020, but the prediction over the Baseline model has the most negative impact on the demand response over each day.

7. Future Directions and Recommendations

This section presents the recommendations and future directions for DR baseline calculations.

7.1. Recommendation and future directions for the AGL C&I customers

From the prediction accuracy results reported in Section 4.2.1, it can be observed that for the following C&I customers, the ML model, namely nonlinear regression (NLR), performs better than the average baseline. Thus, we strongly recommend using NLR for the DR baseline calculation for the following C&I customers:

- Chemical Plant
- Telecom
- Telecom VIC
- Water utility 2
- Water utility 5

However, for other C&I customers (e.g., medium manufacturing, metal recycling, sandstone quarry), we recommend using the average baseline method, since there is no significance difference in the performance with the ML models.

From the results of Section 4.2.2, it can be concluded that using the maximum temperature does not change the model accuracy except for the Shopping Centre profile. Thus, we recommend using the NLR model with daily temperature as a predictor for this baseline calculation.

From the demand response point of view, as reported in Section 4.3, the NRL model shows potentially high monetary benefits for most C&I customers compared to the average baseline calculation, except for Water Utility 1, and Water Utility 2.

The above recommendations are made based on the data available for the C&I customers. If more data is available for testing, the accuracy of the model can be improved. Generally, the common approach is to split the data into 70% for training and 30% for testing. Based on that, this study calculated the testing confidence scores, which is based on the number of testing samples, as shown in Table 7.

Table 7. Machine learning model testing accuracy confidence scores for the AGL dataset.

C&I customer	Data availability		Testing confidence score
	From (dd/m/y)	*To* (dd/m/y)	
Chemical Plant	10/11/2020	14/01/2021	Medium
Medium Manufacturing	22/12/2019	31/01/2020	Low
Metal Recycling	22/12/2019	18/02/2020	Medium
Quarry Sandstone Ops	22/12/2019	31/01/2020	Low
Shopping Centre	1/01/2020	31/01/2020	Low
Telecom	10/11/2020	10/06/2021	Very High
Telecom VIC	1/01/2021	25/05/2021	Very High
University	22/12/2019	31/01/2020	Low
Water Utility 1	1/01/2020	31/01/2020	Very Low
Water Utility 2	22/12/2019	31/01/2020	Low
Water Utility 5	1/01/2020	31/01/2020	Low
Water Utility VIC	1/05/2021	20/05/2021	Not valid

C&I customers with very low and low testing confidence scores, need more data to build more concrete conclusions. C&I customers with testing confidence scores labelled as Not valid means not enough data available for testing. Thus, more data is required for C&I customers with low and very low testing confidence scores to develop more accurate ML models. Also, large time windows are required, since large time windows show better performance, as shown in Section 6.3.

7.2. Recommendation and future directions for the Greater Western Water

For the Greater Western Water dataset, from the results reported in Section 5.2, the machine learning model does not have a significant advantage compared to the conventional average Baseline calculation. Therefore, we recommend using the Baseline calculation for the Greater Western Water dataset. For future work, utilising the Greater Western Water dataset, due to being a long dataset, for the AGL dataset (Water Utility 2, Water Utility 5, Water Utility VIC), could be investigated by using transfer learning. This will overcome the limitation of insufficient testing samples for these C&I customers.

7.3. Recommendation and future directions for the Gippsland Water

The results reported in Sections 6.3.1 and 6.3.2 show that using the previous time stamps to predict future demands can lead to significantly better performance than using the same time stamps. ML methods can provide considerably more accurate predictions than the baselines, especially when sufficient data is available (e.g., the Gippsland Water dataset). We recommend using the previous 10 (i.e., TW = 10) time stamps as the inputs for the Gippsland Water dataset. Also, adding the temperature information can help generate more accurate prediction results, demonstrating that temperature influences demand. Another important finding is that either NLR or LR outperformed the average baseline model, and hence both these ML techniques can be used instead of average baseline model in the future.

Moreover, the energy demand prediction shows that NR and NLR have almost similar performances. From the Appendix, it can be inferred that LR is much simpler and easier technique to implement than the NLR. Therefore, we recommend using the LR as the baseline model, where the temperature information can help improve the future energy demand prediction.

Since the tariff structure is not available for the Gippsland Water dataset, this study has performed some demand response analysis to verify the prediction results' effectiveness, as illustrated in Section 6.3.3. The reported results show that the NLR and LR models' predicted energy demand values are much closer to actual values. Also, a more accurate prediction leads to less difference between the predicted energy demand and the actual demand. The average demand difference percentage of NLR or LR is much less than the average baseline model. In the future, we recommend using the tariff structure to calculate the benefits to further validate these findings from monetary perspective.

8. References

[1] F. Pallonetto, M. De Rosa, F. Milano, and D. P. Finn, “Demand response algorithms for smart-grid ready residential buildings using machine learning models,” *Applied Energy*, vol. 239, pp. 1265–1282, Apr. 2019.

[2] Y. Ozturk, D. Senthilkumar, S. Kumar, and G. Lee, “An Intelligent Home Energy Management System to Improve Demand Response,” *IEEE Transactions on Smart Grid*, vol. 4, no. 2, pp. 694–701, Jun. 2013.

[3] M. D. Hu Ninghao Liu, Xia, “Techniques for Interpretable Machine Learning,” *Communications of the ACM*, vol. 63, no. 1, pp. 68–77, Jan. 2020.

[4] Y. Zhang, W. Chen, R. Xu, and J. Black, “A Cluster-Based Method for Calculating Baselines for Residential Loads,” *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2368–2377, Sep. 2016.

[5] Y. Chen *et al.*, “Short-term electrical load forecasting using the Support Vector Regression (SVR) model to calculate the demand response baseline for office buildings,” *Applied Energy*, vol. 195, pp. 659–670, Jun. 2017.

[6] D. Deltetto, D. Coraci, G. Pinto, M. S. Piscitelli, and A. Capozzoli, “Exploring the Potentialities of Deep Reinforcement Learning for Incentive-Based Demand Response in a Cluster of Small Commercial Buildings,” *Energies*, vol. 14, no. 10, Jan. 2021.

[7] H. Lee, H. Jang, S.-H. Oh, N.-W. Kim, S. Kim, and B.-T. Lee, “Novel Single Group-Based Indirect Customer Baseline Load Calculation Method for Residential Demand Response,” *IEEE Access*, vol. 9, pp. 140881–140895, 2021.

[8] K. Li, F. Wang, Z. Mi, M. Fotuhi-Firuzabad, N. Duić, and T. Wang, “Capacity and output power estimation approach of individual behind-the-meter distributed photovoltaic system for demand response baseline estimation,” *Applied Energy*, vol. 253, p. 113595, Nov. 2019.

[9] EnerNOC, Inc., “The Demand Response Baseline,” EnerNOC, Inc., Boston, USA, White Paper, 2011. [Online]. Available: https://library.cee1.org/sites/default/files/library/10774/CEE_EvalDRBaseline_2011.pdf

[10] “Historical weather observations and statistics, Bureau of Meteorology,” *Bureau of Meteorology Australia*. [Online]. Available: http://www.bom.gov.au/climate/data-services/station-data.shtml (accessed Apr. 30, 2022).

[11] “Annual and monthly heating and cooling degree days - documentation,” *Bureau of Meteorology Australia*. [Online]. Available: http://www.bom.gov.au/climate/map/heating-cooling-degree-days/documentation.shtml (accessed Jun. 30, 2022).

[12] R. J. Mislevy, “Recent Developments in the Factor Analysis of Categorical Variables,” *Journal of Educational Statistics*, vol. 11, no. 1, pp. 3–31.

[13] Australian Energy Regulator, “Annual benchmarking reports 2021,” Australian Energy Regulator, Melbourne, Australia, Benchmarking, Nov. 2021. Accessed: Aug. 30, 2022. [Online]. Available: https://www.aer.gov.au/networks-pipelines/guidelines-schemes-models-reviews/annual-benchmarking-reports-2021.

Appendix: Mathematical models of ML predictors

Linear regression (LR):

$$y = a_1 x_1 + \ldots + a_n x_n + b \tag{A.1}$$

Quadratic polynomial regression (NLR):

$$y = a_1 x_1^2 + \ldots + a_n x_n^2 + b_1 x_1 + \ldots + b_n x_n + c \tag{A.2}$$

$x_1, \ldots, x_n$ represent the inputs for each predicted y. For both LR and NLP, the coefficients $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$, and the constants b and c are obtained via the training process. To evaluate how the predicted y is accurately corresponding to the real value, the prediction result is evaluated via mean absolute error (MAE), root mean squared error (RMSE), defined as follows:

$$\begin{gathered} MAE = \frac{1}{n}\sum_{i=1}^{n}\left| y_i - y_i \right| \\ RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - y_i \right)^2}{n}} \end{gathered} \tag{A.3}$$

where n represents the number of samples. y_i and y_i are the predicted and real values of the i-th sample, respectively.

Support vector regression (SVR):

Another way to represent the model is in the matrix form. For linear regression (LR), the matrix form can be represented as follows:

$$\hat{y} = \boldsymbol{w} \cdot \boldsymbol{x} + b \tag{A.4}$$

where $\boldsymbol{w}$ and b are the parameters of the model to be learned during the training. For SVR, the model tries to find the best set of parameters that reduce the distance between the predicted values and real values within some margin of error ε by minimising the following:

$$\boldsymbol{w} \cdot \boldsymbol{x} + b - \hat{y} \leq \varepsilon \tag{A.5}$$

This distance defines the boundary between the predicated and real values. For a nonlinear boundary, a kernel trick is used, such as Radial basis kernel as follows:

$$Radial\ basis\ kernels\ \ K(x, \hat{y}) = \exp\left(-\frac{1}{2\sigma^2}\left|\left|x - \hat{y}\right|\right|^2\right) \tag{A.6}$$

Bayesian linear regression (BLR):

For BLR, the predicted value in the matrix form is defined as follows

$$\hat{y}_i = \boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{W} + \eta_i \tag{A.7}$$

where η_i is bias value that sampled from the normal distribution $\eta_i \sim N(0, \sigma^2)$.

The BLR assumes that the parameter of model $\boldsymbol{W}$ has a prior distribution as follows:

$$P(\boldsymbol{W})\sim N(M_o, S_o) \tag{A.8}$$

where M_o is the prior mean, and S_o is the prior covariance which is calculated during the training phase. The common approach to calculating M_o and S_o is by using the Maximum Likelihood Estimator (MLE).