DATA SCIENCE FINAL PROJECT
HEX 2 2024

A DATA SCIENCE APPROACH TO PREDICTING ENERGY DEMAND IN NEW SOUTH WALES, AUSTRALIA

Group J

Authors: Siyi Chen(z5347249), Partha Das (z5388973), Kedar Viswanathan (z5345069), Udayan Chelvaratnam (z5043038)

Table of Contents

Quick Links

Github: https://github.com/Aileennini/ds_project_2024

Jupyter Notebooks:

EDA and Statistical Modelling: Group_Project_Report_Group_J.ipynb

ACF and PACF, and ML Modelling: ARIMA_REPORT_Rev01_U.ipynb

Abstract

Energy demand management is an important issue for energy suppliers and policy makers in Australia. While energy consumption is expected to increase overall, the introduction of renewable energy sources has seen demand on traditional energy sources stabilise and even reduce. Australia's national approach to energy management and in particular renewables as response to climate change has often been fragmented and ineffective (Nelson, 2015). As a result, this paper will focus on a specific state, New South Wales (NSW). NSW has seen a 2% decrease in energy consumption from the grid, driven by a 19% increase in renewable energy sources, primarily in the residential sector which only accounts for 11% of overall energy demand in the state (NSW Environment Protection Authority, n.d.). The objective of this paper is to identify the short and long-term influencers of energy demand in NSW and build models to support policy makers and organisations to effectively predict the energy demand to assist with planning for effective and sustainable solutions. We demonstrate that a polynomial regression model can be suitable to predict energy demand in the long term, driven by temperature, solar PV installations, price, GDP, and population. We illustrate that in the long term, solar PV installations have a negative impact on aggregate energy demand, while temperature and price are the biggest influencers of demand. In the short term, an ARIMA[1] model can go to certain length to predict energy demand albeit with limitations, specifically at the extrema. Policymakers and governments may therefore consider further extension of policies and incentives driving greater adoption of solar panels, and in turn, take off significant amount of load from traditional sources of energy that is supporting the growing demand. An ARIMA model may also enable energy suppliers to smooth electricity demand and ensure greater stability of the energy supply. Periods of high demand can be smoothed out by limiting identifying those extrema and limiting energy drain during these periods, by incentivising energy efficient appliances that offset energy demand.

In either situation, an accurate prediction is imperative to sustainable demand management and meeting climate change obligations of the energy industry as well the country.

Introduction

With the advent of climate change, rapid adoption of technology, growth in population and life expectancy, and continued industrialisation and globalisation in the twentieth century, energy demand has increased exponentially during the last century, and the trend is continuing. Global energy consumption is expected to increase by 50% within this decade if the current consumption pattern continues (Smith et al., 2007). Disruptions in power can have serious economic impacts, both for industries, and individual companies. In December of 2023, the Australian Energy Market Commission (AEMC) introduced new rules for National Electricity Market to ensure energy reliability as the country transitions to increased reliance on renewable energy sources (Latief, 2023). These rules include incremental raises to the Market Price Cap, Cumulative Price Threshold and Administrative Price Cap with the aim to provide greater flexibility for investors to contribute to new generation and storage infrastructure for periods of high demand.

Energy demand management is an important issue for energy suppliers and policy makers in Australia. While energy consumption is expected to increase overall, the introduction of renewable energy sources has seen demand on traditional energy sources stabilise and even

[1] Autoregressive Integrated Moving Average

reduce. By 2030, the Australian Energy Market Operator (AEMO) expects 50% of consumers to be driving demand to the National Electricity Market to meet their energy needs (Energy Security Board, 2021). Two of the factors that will significantly impact energy demand in the next decade is Climate change which will impact average and extreme Temperatures across Australia (Ahmed, Muttaqi and Agalgaonkar, 2012), and dedicated investment in Electric Vehicle (EV) infrastructure (Zhang et al., 2017) and the subsequent growth of EVs.

However, Australia's national approach to energy management and in particular renewables as response to climate change is fragmented and ineffective (Nelson, 2015). As a result, this paper will focus on a specific state, New South Wales (NSW). NSW has seen a 2% decrease in energy consumption from the grid, driven by a 19% increase in renewable energy sources, primarily in the residential sector which only accounts for 11% of overall energy demand in the state (NSW Environment Protection Authority, n.d.). While this is exciting, the NSW Government is planning to invest $209m in an EV charging network, when coupled with additional policies and incentives is intended to make Evs represent 52% of new car sales in NSW by 2031 (NSW Climate and Energy Action, 2022), which may offset some of this demand stabilisation.

With competing pressures to tackle climate change, it is not a viable option to meet this demand through traditional energy sources. This presents a significant challenge for NSW, as current strategies and lack of political momentum has resulted in limited success in finding other sources of energy through large scale renewable energy assets. The objective of this paper is to identify the short and long-term influencers of energy demand in NSW and build models to support policy makers and organisations to effectively predict the energy demand so that they can plan for the right solutions. This paper will use a Regression model to predict long term changes in energy demand, and an ARIMA model to show the short-term prediction.

Literature Review

Australian energy market overview

International Energy Agency (IEA) publishes reports and statistics on energy usage for various countries, covering about 80% of global energy usage (International Energy Agency, n.d.). As per statistics published by IEA (IEA, 2019), total electricity consumption in Australia was about 253 TWh in 2022. Figure 1 shows that there has been slight decline in industrial electricity consumption since 2010 but residential electricity consumption has continued to increase following a slump in 2013-14. The decline in industrial electricity consumption can be attributed to reduced growth of larger industrial energy users, closures of industrial facilities, energy efficient programs and embedded generation. The reduction in residential electricity consumption during 2013-14 may be attributed to energy efficiency programs, widespread deployment of rooftop solar PV, water heating fuel-switch program and consumer response to increasing retail electricity price (Sandiford et al., 2015).

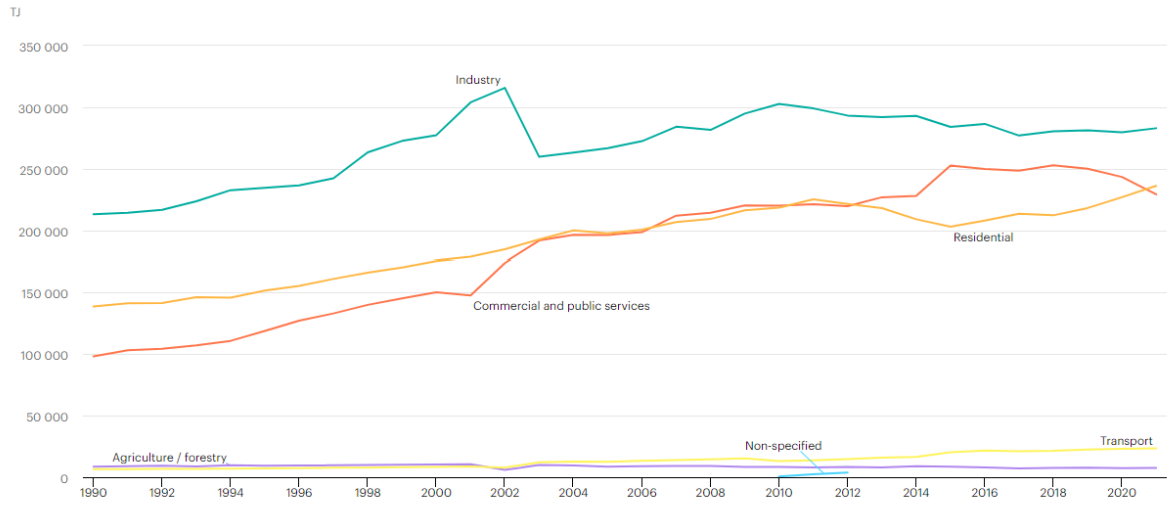


Figure 1: Energy consumption by sector, Australia, 1990-2021 (IEA, 2019)

In figure 2 it is noticeable that per capita electricity usage has reduced during the last decade, following the peak of 11 MWh per capita consumption during early 21st century.

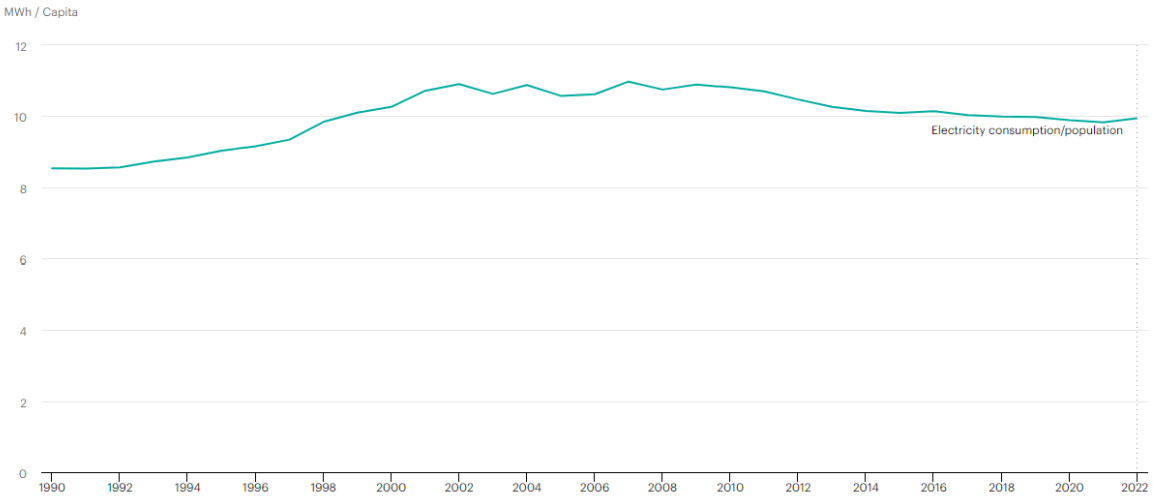


Figure 2: Energy consumption per capita, Australia, 1990-2022 (IEA, 2019)

While in Australia more than 50% of electricity is generated still being generated by traditional sources, it is noticeable that wind and solar PV account for more than 20% of the total electricity generated in 2022.

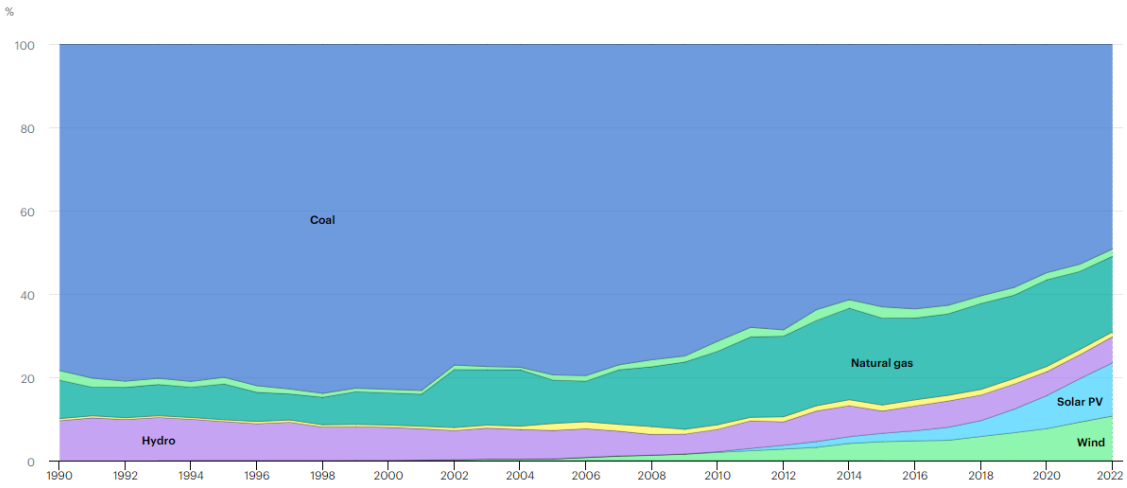


Figure 3: Distribution of energy generated by source, Australia, 1990-2022 (IEA, 2019)

Australian Energy Market Outlook

A look into Australia's energy strategies and frameworks is fundamental to understand the shift in energy source, and how the shift may evolve in the future. The Australian Government's Powering Australia plan (Department of Climate Change, Energy, the Environment and Water, 2023b) is focused on lowering emissions by boosting renewable energy and reducing pressure on energy bills. Of particular note is the National Energy Transformation Partnership (Energy.gov.au, 2022) established in 2022. It is a framework for Commonwealth, state and territory governments to work together on reforms to help transform Australia's energy systems to achieve net zero target by 2050. Some of the key themes of this partnership are:

- Planning for adequate energy generation and storage
- Understanding demand evolution
- Coordinating gas and electricity planning
- Enhancing energy security management and
- Accelerating nationally significant transmission projects

In addition to above, the Australian Government's Department of Climate Change, Energy, the Environment and Water (DCCEEW) recently published the National Energy Performance Strategy (Department of Climate Change, Energy, the Environment and Water, 2023a) which provides a long-term framework to manage energy demand. There are 5 focus areas in this strategy.

- Economy wide momentum
- Households
- Communities, businesses and industries
- Energy system
- Technology and innovation

The focus of energy system, in the National Energy Performance Strategy, is to drive Australia's energy transformation. The government is strengthening the role of demand side in energy system planning and governance frameworks to support ongoing energy performance measures.

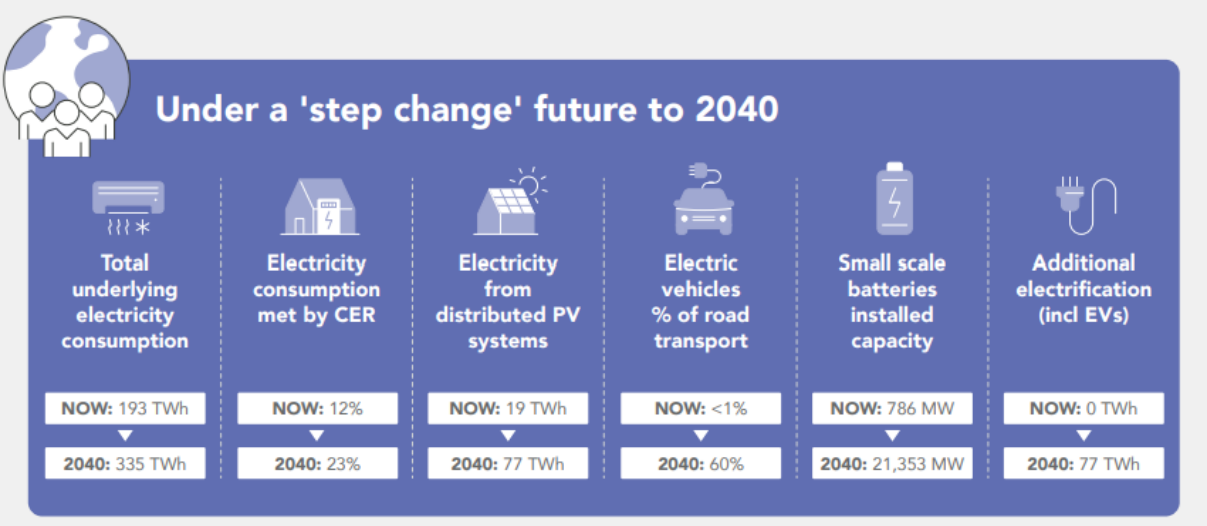
The Australian Energy Market Operator (AEMO) develops energy planning scenarios used in forecasting and planning analysis and publications for the National Electricity Market (NEM). (Australian Energy Market Operator, 2023a). The 2023 Electricity Statement of Opportunities (OSOO) assumes three scenarios to forecast energy demand in Australia. (Australian Energy Market Operator, 2023b)

1. **Green energy exports:** A scenario where rapid transformation happens in the energy sector, including strong use of electrification, green hydrogen and biomethane.

2. **Step change (ESOO central scenario**): A central scenario where the scale of energy transformation supports Australia's contribution to limiting global temperature rise to below 2degree C compared to pre-industrial levels. It relies on a strong contribution from orchestrated consumer energy resources (CER), strong transport electrification, and opportunities for Australia's larger industries to electrify to reduce emissions, or to use developing hydrogen production opportunities or other low emissions alternatives to support domestic industrial loads. This is considered to be the most likely scenario.

3. **Progressive change:** A scenario where transformation is sufficient to meet Australia's commitment to 43% emissions reduction by 2030 and net zero emissions by 2050, but under challenging economic conditions and higher relative technology costs.

Under the step change scenario, total underlying electricity consumption is expected to increase further by 2040. However, electricity consumption met by consumer energy resources is forecasted to be almost double during this period, from 12% to 23%, and electricity generated from distributed PV systems is forecasted to increase by 300%, from 19 TWh to 77 TWh.



Source: (Australian Energy Market Operator, 2023a)

The central forecast scenario shows that underlying electricity consumption will continue to increase, driven by population growth, economic activity, and emerging opportunities to electrify new customer. At the same time, CER and energy efficiency is expected to slow operational consumption growth. The figure below with actual and forecast under the central scenario show that while underlying residential and business electricity consumption is expected to grow, reductions from rooftop PV and energy efficiency will meet significant part of the increased demand.

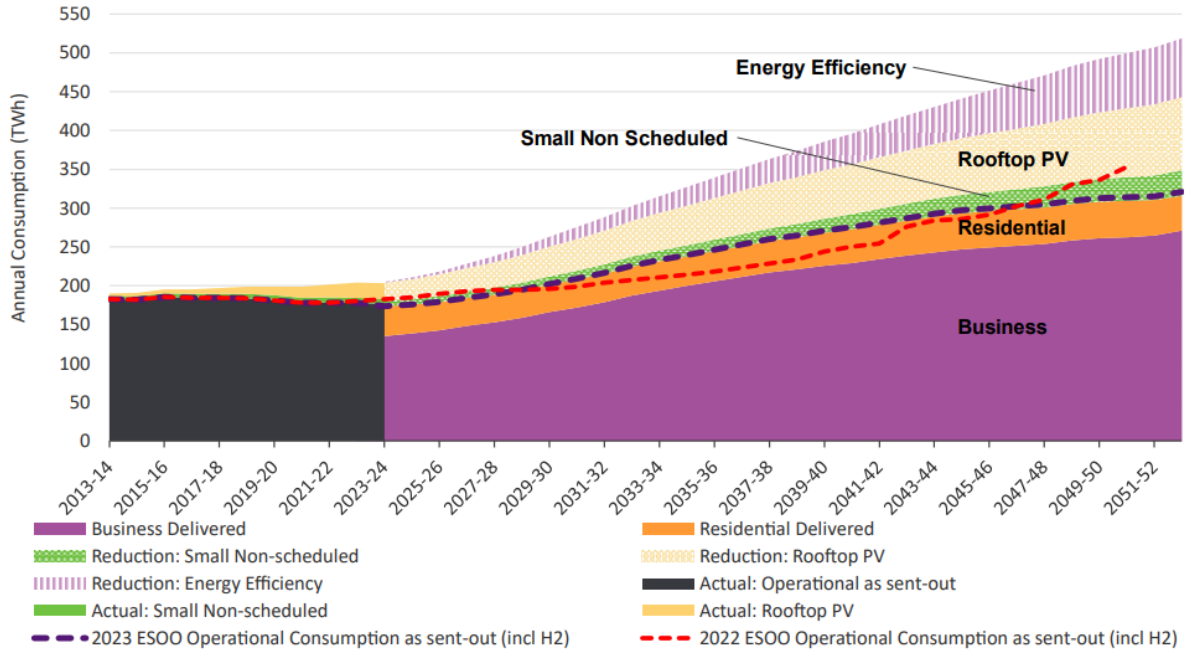


Figure 4: Actual and forecast NEM electricity consumption, ESOO central scenario

The underlying residential electricity demand is expected to increase significantly. Under the 2023 ESOO Central scenario, underlying residential consumption is forecast to increase 31% over the next decade, from approximately 57 TWh in 2022-23 to 75 TWh in 2032-33. This is primarily due to transition from traditional vehicles to electric vehicles (EVs) and growth in residential dwellings (driven by population growth). The model assumes that by 2032-33, approximately 30% of residential passenger vehicles or more than 4 million residential passenger vehicles will be EVs, consuming 11 TWh per annum. Construction of around 1.7 million new dwellings is forecast to increase consumption by 12 TWh per annum, along with electrification of space heating, hot water heating and switch from gas cooking appliances to electricity. However, distributed PV generation is also increasing simultaneously. It is expected that within the next decade, roughly half of residential sectors underlying consumption will be fulfilled by distributed PVs in the NEM, which is forecasted to grow to 60% in the long term and thereby keeping the underlying residential consumption stable over the longer term.

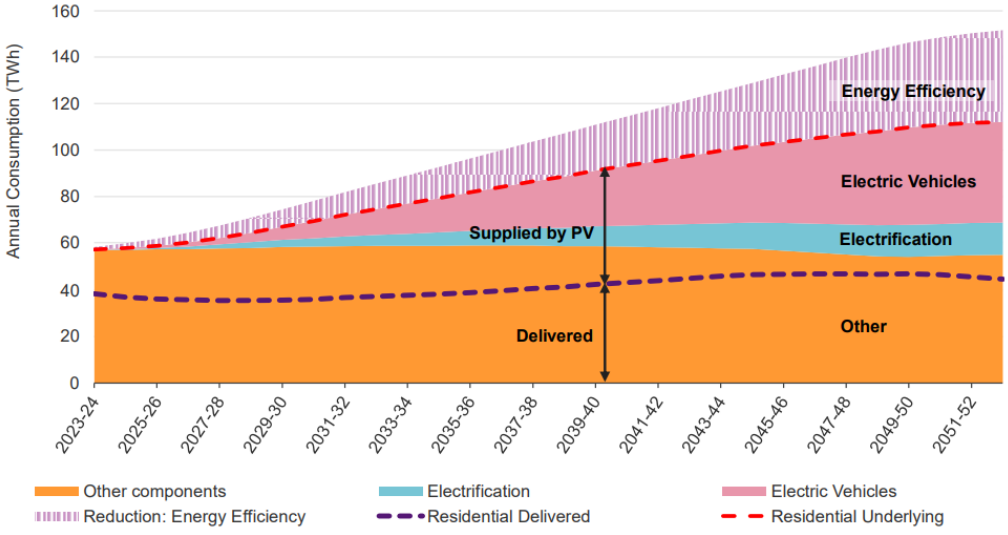


Figure 5: Forecast of residential consumption, ESOO central scenario

Industrial consumption is driven by economic conditions, global commodities market, and emerging hydrogen production. In contrast with residential consumption, operational consumption is expected to grow for industrial sector. While energy efficiency gains are anticipated to increase in future, it is not sufficient to meet the increasing demand for business. Therefore, both underlying and operational electricity consumption for businesses are expected to continue to rise.

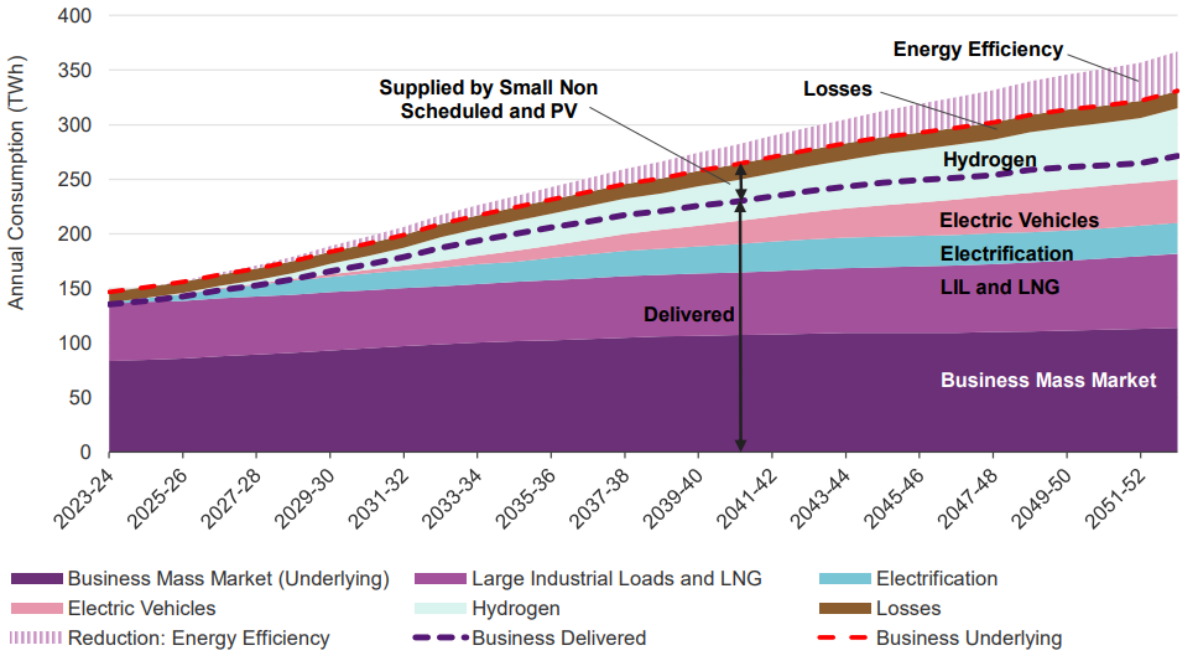


Figure 6: Forecast of business consumption, ESOO central scenario

Significance and challenges of forecasting

Energy demand forecasting is essential to ensure sustainable economic growth and environmental security. It plays a critical role in energy supply-demand management by private suppliers as well as government agencies (Islam et al., 2020). While a private supplier may be interested in accurate forecast to meet customer demand, set appropriate price and meet investors' expectations, government agencies are interested in ensuring suitable allocation of energy resources, deciding upon construction of infrastructure, framing policies to meet and/or influence future demand and drafting strategies for reduced emission. Energy demand management considers a series of technical, organisational, and behavioural solutions to decrease consumer demand. A range of cost effective and environmentally friendly yet commercially viable solutions are being explored. Accurate forecast of demand is expected to promote change in consumer usage pattern, cost effectiveness and ultimately achieve self-sufficiency.

Energy demand models have been studies by many researchers, in many countries across the world and with varying perspectives. The models are usually developed specific to a nation or utility, depending on economic and market conditions prevailing at the time of prediction and expected evolution in those conditions (Islam et al., 2020). The models can be classified in several ways – static versus dynamic, univariate versus multivariate and techniques like conventional time series, engineering models and hybrid models.

Due to availability of real measurements of historical demand data, statistical parametric models are often used to describe and forecast energy demand, particularly for residential consumption. A study (Verdejo et al., 2017) summarised a comparison of such techniques for residential energy demand forecast models as below.

Category	Advantage	Disadvantage
Linear	Easy implementation and interpretation. Estimation techniques well established. Scientific acceptance and a wide range of applicability.	In real world a few phenomena correspond with models assumptions, this leads sometimes, does not always provide useful results.
Non-linear	More general than linear methods, which gives a major flexibility at the moment to fit a data series.	The function that gives the optimum fit should be determined, this hinders the preparation analysis. Less amount of validations tools: for example, there is not exist a explicit calculus for the R2 coefficient.
Discrete	Uses and gives more simplified information	Analysis and result too robust for short time intervals studies.
Continuous	Wide application field in the description of any phenomenon	Estimation, simulation and validation techniques more complex and sophisticated. For good parametric estimation, a lot data and/or small time intervals between observation are required. Explicit analytic solutions do not always be able and numeric approximations must to be used exist
Parametric	Greater provision of information, due to certain probability distribution is assumed for the data.	To assume that the data come from of a specific probabilistic model, biased conclusions could be obtained if a wrong model is used.
Non-parametric	Less condition about the data should be assumed. Which is better in situations when the truly distribution is unknown or cannot be approximated easily.	Limited software implementation. Oriented to hypothesis test, instead of effect estimations. Non-parametric estimations and confidence intervals extraction does not easy.

Table 1: Modelling techniques comparison for residential energy demand (Verdejo et al., 2017)

In another study (McLoughlin, Duffy and Conlon, 2013) time series approaches to forecast demand at individual dwelling level were evaluated. The study used individual and aggregate demand data from the Irish Transmission System Operator, Eigrid. Analysis of aggregated data demonstrate that typically there are consistent peaks in the morning, lunch, and evening

times. The shape varies marginally over the course of the year due to seasonality. However, for a single consumer, the profile shape can change significantly from one day to the next. Similarly, the demand profile can vary significantly from one consumer to another, on the same day. Such variations pose a challenge for accurate forecasting in the short term.

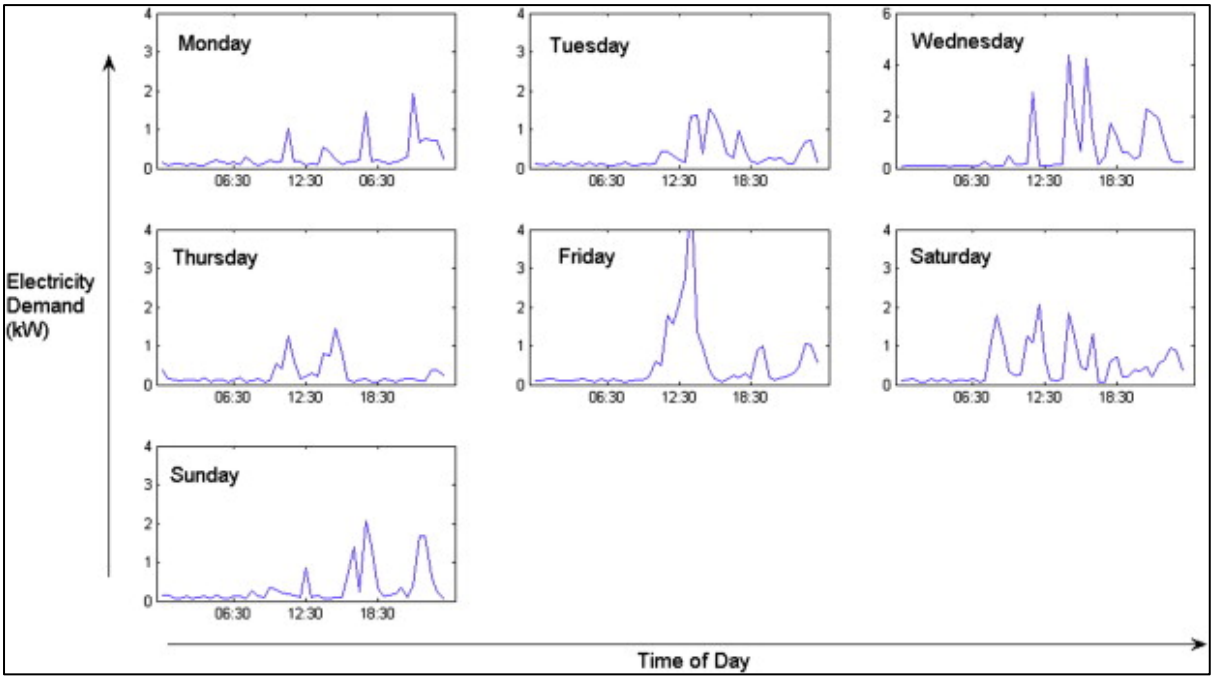


Figure 7: Daily electricity load profiles for a single randomly chosen customer over a weekly period showing intra-daily variations (McLoughlin, Duffy and Conlon, 2013)

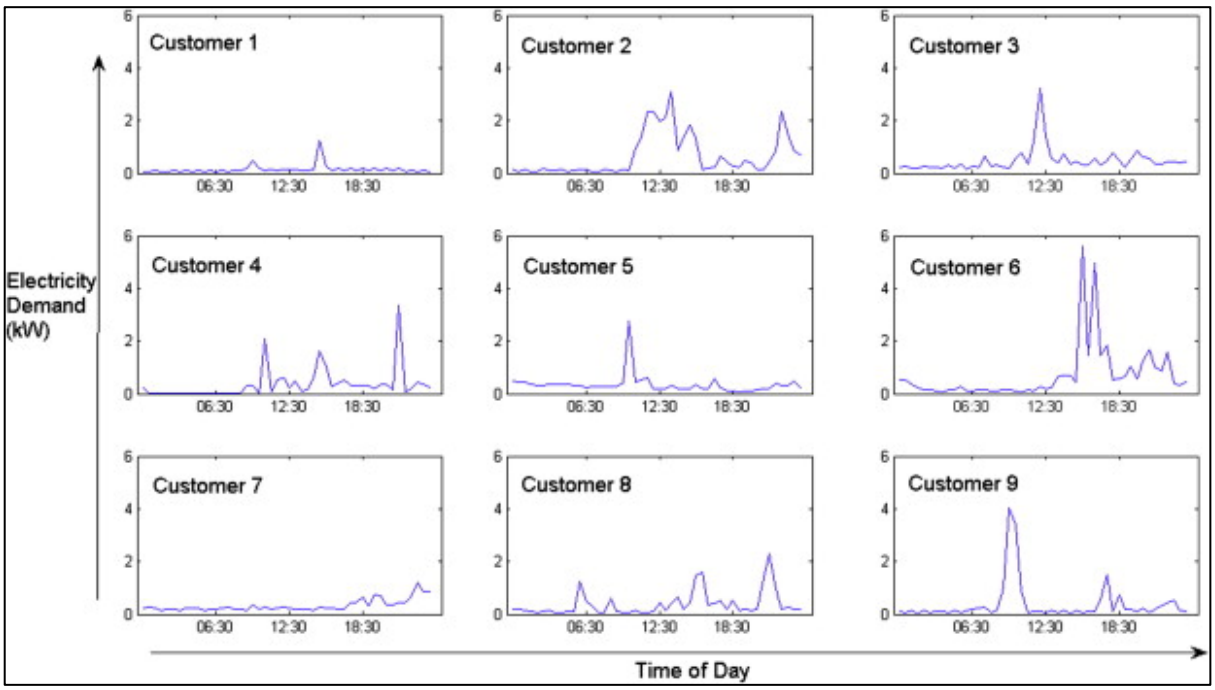


Figure 8: Daily electricity load profiles for nine randomly chosen customers illustrating variation between dwellings (McLoughlin, Duffy and Conlon, 2013)

It was found that neural networks are good at characterising highly non-linear nature of residential electricity demand profile, which can have many sharp changes in demand within a small interval. The model can perform particularly well in shorter time intervals. A Gaussian

process is good at approximating small intervals of sharp electricity demand, but it was found to be less good at approximating smoother average demand. A multiple regression model is widely used for standard forecasting. In order to replicate the variability of electricity load profile, the model needs to characterise each half-hour period separately. It is the method of choice for the UK grid operator, National Grid, to develop standard load profiles for the purposes of electricity settlement. Autoregressive models are also widely used to profile aggregate electricity demand (as opposed to individual demand profile). Finally, the study found that Fourier transformation had the ability to characterise temporal and magnitude components of demand profile. Scalability is additional benefit of this transformation, although the model had difficulty characterising small intervals of high demand. Fourier transforms and Gaussian processes showed the greatest potential for characterising domestic electricity demand load profiles.

While short-term (an hour to a week) or medium-term (a month to 5 months) forecasting is used for planning energy production, tariffs etc., long-term forecasting (5-20 years) is applied for resource management, strategic planning and investment in infrastructure (Ghalehkhondabi et al., 2016). Majority of forecasting horizons are shorter in duration – hourly, daily, weekly, or monthly. It is assumed that the time series formed by energy consumption data can be accumulated accurately, and previous values can be used to forecast over shorter time horizons. In such forecasting, input variables capture short term variabilities, such as temperature and humidity which vary hourly or half-hourly and are recorded accurately. On the other hand, energy demand changes with time, climatic variables, socioeconomic and demographic parameters which poses difficulties for accurate long-term forecasting. Socioeconomic and demographic parameters, such as Gross Domestic Product (GDP), population of an area considering long-term forecasting, increase in number of dwellings and changes in types of dwellings etc. are measured annually and can be used for long-term forecasting.

Short term forecasting

For short term forecasting purpose, factors such as temperature, humidity, population density along with past consumption trend is more suitable (Hagan and Behr, 1987) (Nogales et al., 2002). ARIMA or SARIMA models are widely used to model non-stationary time series which contains trend or seasonality. Kareem and Majeed (Kareem and Majeed, 2006) used the SARIMA model to forecast the monthly peak load demand for the Sulaimany Governorate in Iraq. They proposed a SARIMA model and evaluated it based on measuring the MAPE from data created during the year 2005. The ARIMA and Generalized Autoregressive Conditional Heteroscedasticity (GARCH) models were used by Hor et al. (Hor, Watson and Majithia, 2006) to forecast the daily electricity consumption. Sigauke and Chikobvu (Sigauke and Chikobvu, 2011) used the different combinations of regression, SARIMA and GARCH models to forecast the daily peak electricity demand in South Africa. Results of developed models are compared with a piecewise linear regression model and approved that the developed model including all regression, SARIMA and GARCH models has the best accuracy among all other individual and combined methods.

A study to forecast peak daily electricity demand in NSW, Australia used half hourly demand data and an ARIMA (autoregressive moving average) model (As'ad, 2012). The analysis demonstrated that ARIMA model based on past 3 months data is likely to perform best (compared to six-, nine- or twelve-months data) to forecast short-term demand, based on RMSE (root mean squared error) and MAPE mean absolute percentage error) measures of accuracy.

Long-term forecasting

A multiple linear regression model for annual electricity consumption in New Zealand used GDP, average price of electricity and population and found strong correlation between the predictors and electricity consumption (Mohamed and Bodger, 2005).

Energy supply and demand for the Asia-Pacific region is analyzed using econometric factors (GDP, foreign trade) with oil prices, domestic oil prices, and substitution (Intarapravich et al., 1996). An econometric model is defined by the following steps: developing economic hypothesis; a mathematical model of the hypothesis; an econometric model of the hypothesis; an estimation of the econometric model; testing the hypothesis; and forecasting (Kayacan et al., 2012).

System Dynamic model, a computer-oriented mathematical modelling approach that uses inter-relation of variables in a complex setting including time-to-time variation in system behaviour and a feedback loop to consider new system conditions, has been used to forecast urban energy consumption trends under different growth scenarios. The model was divided into sub-models for residential, commercial, industrial and transportation (Fong et al., 2007). In various other such implementations, following predictors were considered:

- Regional services, population, regional attractiveness etc. (Vaudreuil, M.P.,2011)
- Per capita consumption of electricity and population (Akhwanzada and Tahar, 2012)

Methods, Tools and Data

Methods

From a modelling perspective we wanted to look at both long-term and short-term models to predict the energy demand. There are multiple methods to be used for this problem, all of which have pros and cons. We used the following ones in our modelling after researching the data:

Polynomial Regression

Polynomial regression is a widely used statistical method for modelling non-linear relationships between variables. Compared to other models, it has following advantages:

- Simplicity and Interpretability: Polynomial regression models are straightforward to understand and interpret. They provide clear insights into how each predictor variable impacts the target variable. This simplicity makes it easy to understand model outputs and decisions based on these outputs.
- Efficiency: Polynomial regression can be computationally less intensive compared to more complex models. This makes it suitable for situations where computational resources are limited.
- Speed: Due to its simplicity, polynomial regression models can be trained very quickly.
- Less Data Required: Polynomial regression can perform well with a relatively small amount of data and does not require as much data as more complex models to provide useful insights.

ARIMA (Autoregressive Integrated Moving Average)

ARIMA is a popular statistical method for forecasting non-stationary time series data. This model incorporates both autoregressive (AR) and moving average (MA) components, along with a differencing pre-processing step to make the series stationary. Here are the advantages of using ARIMA modelling:

- Flexibility: ARIMA models can be configured in numerous ways to adapt to the specific characteristics of a time series, such as the presence of trends or seasonal patterns. By adjusting its parameters (p, d, q) and seasonal components (P, D, Q, s), ARIMA can be tailored to fit a wide range of time series.
- Predictive Power: When properly fitted, ARIMA models can provide highly accurate forecasts for time series data, assuming that future patterns and trends will follow those in the past.
- Statistical Foundation: ARIMA is based on solid statistical foundations, providing confidence intervals and significance tests for its forecasts, which are valuable for understanding the reliability of predictions.

Tools

There were a number of tools that we used over the course of the analysis and paper.

Project Management

From an overall project management perspective, we predominately used Microsoft Teams to collaborate and communicate. We also supplemented this, where required, with collaboration tools like Zoom, and Monday.com to keep track of tasks to make sure that we were all accountable for our contributions.

Report and Presentation

For the report and presentation, we used:

- Jupyter Notebook: we used Jupyter Notebook to build out the code and report in a consumable and collaborative format.
- Word: to build out the report in a consumable format and remove the bulk of code and results that were not required.
- PowerPoint: to build the presentation and record.

Analysis and Modelling

For the analysis and modelling, we used the following tools:

- Github: GitHub is a platform for version control and collaboration. It allows our team to work together on projects from anywhere, managing the code in repositories. We used it to track changes in software development projects, manage code histories, and facilitate collaboration among multiple contributors through features like issues, pull requests, and code reviews.
- Python: we used the Pandas and NumPy libraries for the data analysis. We felt between Python and R, the Python language was best suited to our skillset.
- Statsmodel: Statsmodels is a Python module that provides classes and functions for the estimation of many different statistical models, as well as for conducting statistical tests and statistical data exploration. We used the time-series analysis model ARIMA to do the modelling.
- Scikit-Learn: was used in both the Statistical and Machine Learning models that we looked at

Below is a code snippet that shows the full tools that were used for the analysis:

```
# Import necessary libraries
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_absolute_error, mean_absolute_percentage_error, mean_squared_error
from statsmodels.tsa.stattools import adfuller
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
from statsmodels.tsa.arima.model import ARIMA
```

Data

For our analysis and modelling, we relied on six data sources. These files can be found in the "data" folder of the linked Github repository.

1. **Temperature Data**: Market Management System database, which is published by the market operator from the National Electricity Market (NEM) system. (https://github.com/Aileennini/ds_project_2024/tree/main/data/NSW).

2. **Energy Demand for NSW**: Market Management System database, which is published by the market operator from the National Electricity Market (NEM) system. (https://github.com/Aileennini/ds_project_2024/tree/main/data/NSW).

3. **Population for NSW:** Australian Bureau of Statistics - https://www.abs.gov.au/statistics/people/population/national-state-and-territory-population/dec-2022).

4. **Australian GDP**: Australian Bureau of Statistics - https://www.abs.gov.au/statistics/economy/national-accounts/australian-national-accounts-national-income-expenditure-and-product/latest-release#state-and-territory-final-demand).

5. **Electricity Prices for NSW**: CEIC - https://www.ceicdata.com/en/australia/electricity-prices/electricity-average-peak-price-new-south-wales.

6. **Solar Installations for NSW**: Clean Energy Regulator, Australian Government - https://cer.gov.au/markets/reports-and-data/small-scale-installation-postcode-data#installation-numbers-for-small-scale-systems-by-state/territory.

These datasets encompass crucial variables for our analysis, namely temperature, population, GDP, and energy demand.

Data Cleaning

The only data that needed cleaning was the Temperature data. When looking into the temperature data, we see that for most of the data, the granularity is 30 mins. However there are some cases that the time does not fall into that range. So we needed to process the data to make it more clean. However, we can see after cleaning, the temperate data is still not continuous.

```
df_temp['time_diff_minutes'].unique()
array([ nan,  30.,  60.,  90., 300., 150., 600., 810., 120.,
       180., 330., 240., 690.,   0., 210., 1050., 5430., 450.])
```

Temperature

This cleaned dataset has 3 columns: location, datetime and temperature. "location" is always Bankstown so we don't need to include it in our analysis or modelling. "datetime" has a granularity of 30mins, and that is the granularity in our short-term models. "temperature" column contains the temperature column. From the dataset, it is reasonable to believe the unit is Celsius degree.

temperaturecount	195947
mean	17.530995
std	5.884212
min	-1.3
25%	13.5
50%	17.9
75%	21.5
max	44.7

Table 2: Temperature Data Spread

Here is a sample of the data:

	Location	Date/Time	Temperature (Celsius)
0	Bankstown	1/1/2010 0:00	23.1
1	Bankstown	1/1/2010 0:01	23.1
2	Bankstown	1/1/2010 0:30	22.9
3	Bankstown	1/1/2010 0:50	22.7
4	Bankstown	1/1/2010 1:00	22.6

Table 3: Temperature Data Sample

Energy Demand

This dataset has 3 columns: datetime, total demand, regionid. "datetime" has a graduality of 30 mins, the same as temperature_nsw.csv. "totoldemand" is the Y value we want to predict. "regionid" is always "NSW1" and we do not need to include it into our models.

Here is a sample of the data:

	Date/Time	Total Demand (MW)	Region ID
0	01/1/2010 0:00	8038.00	NSW 1
1	1/1/2010 0:30	7809.31	NSW 1
2	1/1/2010 1:00	7483.69	NSW 1
3	1/1/2010 1:30	7117.23	NSW 1

4	1/1/2010 2:00	6812.03	NSW 1

Table 4: Energy Demand Data Sample

Population

This dataset has 2 columns: time and population. "time" has a granularity of 3 months, which is quite different from the dataset of temperature_nse.csv or totaldemand_nsw.csv. Even though the time granularity of this dataset is different, we still want to include it into out modelling, because it is reasonable to believe there is a correlation between population and totaldemand. For consistency, we can transform the granularity of the totaldemand to 3 months and create new Y values.

Here is a sample of the data:

	Time	Population
0	Dec-2009	7,101,504
1	Mar-2010	7,128,356
2	Jun-2010	7,144,292
3	Sep-2010	7,162,726
4	Dec-2010	7,179,891

Table 5: Population Data Sample

Australian GDP (GDP.csv)

The GDP data reflects the national figure rather than being specific to New South Wales (NSW), as such detailed state-level data was not accessible. This dataset has two columns: time and GDP. Same as population_nsw.csv, the granularity of time is also 3 months. We could only find GDP at the national level but not NSW specifically. Despite the limitation, high economical activity, leading to higher GDP, is expected to be follow higher industrial activity and therefore higher energy demand of industrial sector. We can assume NSW consistently contribute a relatively stable portion to Australian GDP and include this variable into our modelling.

Here is a sample of the data:

	Time	GDP (AUD M
0	Dec-2009	334,934
1	Mar-2010	314,838
2	Jun-2010	340,575
3	Sep-2010	345,512
4	Dec-2010	365.403

Table 6: GDP Data Sample

Electricity Prices (electricity_price_nsw.csv)

This dataset has 3 columns: year, region, avgrrp. "year" has a granuality of 1 year, which is different from the above 30mins or 3 months. So to inlcude this in our long-term modeling, we can assume the change is linear within each year. In this way we can convert the yearly data to quarterly data. "region" is always "NSW". "avgrrp" is the price.

Here is a sample of the data:

	Year	Region	Average Price (AUD/year)
0	2010	NSW	44.19
1	2011	NSW	36.74
2	2012	NSW	29.67
3	2013	NSW	55.10
4	2014	NSW	52.26

Table 7: Price Data Sample

Solar PV Installations (home_solar_nsw.csv)

This dataset has 3 columns: year, nsw, solar_install. "nsw" is the yearly installation of small-scale solar panels in NSW. And "solar_install" is the accumulated solar panel installation. We can also convert this yearly data to quarterly data for modelling.

Here is a sample of the data:

	Year	Solar installations (incremental)	Solar Installations (cumulative)
0	2009	14,008	14,008
1	2010	69,988	83,996
2	2011	80,272	164,268
3	2012	53,961	218,229
4	2013	33,998	252,227

Table 8: Solar Installations Data Sample

Data Preparation

The data we were able to source, did not have the key features, in a consistent manner, and in the time periods we needed it, (i.e. before the end of the quarter we want to predict). To prepare the data suitable for analysis, both training and testing a model, we needed to modify our dataset. Below is sample code of how we modified the data:

Population

```
df_population_quarterly = df_population.copy(deep=True)

# offset the time by 3 months
df_population_quarterly['time'] = df_population['time'] + pd.DateOffset(months=3)
df_population_quarterly['time'] = df_population_quarterly['time'] + pd.offsets.MonthEnd(0)
```

GDP

```
df_gdp_quarterly = df_gdp.copy(deep=True)

# offset the time by 3 months
df_gdp_quarterly['time'] = df_gdp['time'] + pd.DateOffset(months=3)
df_gdp_quarterly['time'] = df_gdp_quarterly['time'] + pd.offsets.MonthEnd(0)
```

Solar Installations

We only have annual data for solar Installations. To make the data usable for our model we can convert yearly solar installation data to quarterly data using linear interpolation within each year.

```
quarterly_solar_install = {'year_quarter': [], 'solar_install': []}

for i in range(len(df_solar_install)):
    if i == 0:
        growth = (df_solar_install.loc[i + 1, 'solar_install'] - df_solar_install.loc[i, 'solar_install']) / 4
        for q in range(1, 5):
            quarterly_solar_install['year_quarter'].append(f"{df_solar_install.loc[i, 'year']}-Q{q}")
            quarterly_solar_install['solar_install'].append(int(df_solar_install.loc[i, 'solar_install'] + growth * (q -
1)))
    else:
        growth = (df_solar_install.loc[i, 'solar_install'] - df_solar_install.loc[i - 1, 'solar_install']) / 4
        for q in range(1, 5):
            quarterly_solar_install['year_quarter'].append(f"{df_solar_install.loc[i, 'year']}-Q{q}")
            quarterly_solar_install = {'year_quarter': [], 'solar_install': []}

for i in range(len(df_solar_install)):
    if i == 0:
        growth = (df_solar_install.loc[i + 1, 'solar_install'] - df_solar_install.loc[i, 'solar_install']) / 4
        for q in range(1, 5):
            quarterly_solar_install['year_quarter'].append(f"{df_solar_install.loc[i, 'year']}-Q{q}")
            quarterly_solar_install['solar_install'].append(int(df_solar_install.loc[i, 'solar_install'] + growth * (q - 1)))
    else:
        growth = (df_solar_install.loc[i, 'solar_install'] - df_solar_install.loc[i - 1, 'solar_install']) / 4
        for q in range(1, 5):
            quarterly_solar_install['year_quarter'].append(f"{df_solar_install.loc[i, 'year']}-Q{q}")
            quarterly_solar_install['solar_install'].append(int(df_solar_install.loc[i, 'solar_install'] + growth * (q - 1)))

df_quarterly_solar_install = pd.DataFrame(quarterly_solar_install)

def quarter_to_date(quarter_str):
    year, q = quarter_str.split('-')
    if q == 'Q1':
        return f'{year}-03-31'
    elif q == 'Q2':
        return f'{year}-06-30'
    elif q == 'Q3':
        return f'{year}-09-30'
    elif q == 'Q4':
        return f'{year}-12-31'
df_quarterly_solar_install['time'] = df_quarterly_solar_install['year_quarter'].apply(quarter_to_date)

# Convert the 'end_of_quarter' column to datetime
df_quarterly_solar_install['time'] = pd.to_datetime(df_quarterly_solar_install['time'])

# Offset the time by 1 year
df_quarterly_solar_install['time'] = df_quarterly_solar_install['time'] + pd.DateOffset(months=12)
```

Pricing

We had to adjust the pricing data in a similar way to the solar Installations.

```
# Same as what we did to solar install, convert price to quartely data
quarterly_price = {'year_quarter': [], 'avgrrp': []}

for i in range(len(df_electricity_price)):
    if i == 0:
        growth = (df_electricity_price.loc[i + 1, 'avgrrp'] - df_electricity_price.loc[i, 'avgrrp']) / 4
        for q in range(1, 5):
            quarterly_price['year_quarter'].append(f"{df_electricity_price.loc[i, 'year']}-Q{q}")
            quarterly_price['avgrrp'].append(df_electricity_price.loc[i, 'avgrrp'] + growth * (q - 1))
```

```
    else:
        growth = (df_electricity_price.loc[i, 'avgrrp'] - df_electricity_price.loc[i - 1, 'avgrrp']) / 4
        for q in range(1, 5):
            quarterly_price['year_quarter'].append(f"{df_electricity_price.loc[i, 'year']}-Q{q}")
            quarterly_price['avgrrp'].append(df_electricity_price.loc[i, 'avgrrp'] + growth * (q - 1))

df_quarterly_electricity_price = pd.DataFrame(quarterly_price)
df_quarterly_electricity_price['time'] = df_quarterly_electricity_price['year_quarter'].apply(quarter_to_date)

# Convert the 'end_of_quarter' column to datetime
df_quarterly_electricity_price['time'] = pd.to_datetime(df_quarterly_electricity_price['time'])
# Offset the time by 1 year
df_quarterly_electricity_price['time'] = df_quarterly_electricity_price['time'] + pd.DateOffset(months=12)
```

Temperature

We had to modify the temperature to be reflective of a quarterly period as well:

```
df_temp['time'] = pd.to_datetime(df_temp['datetime'], format='%d/%m/%Y %H:%M')
df_temp.set_index('time', inplace=True)
temp_quarterly = df_temp['temperature'].resample('Q').mean()
df_temp_quarterly = temp_quarterly.reset_index()
df_temp_quarterly['time'] = df_temp_quarterly['time'] + pd.DateOffset(months=3)
df_temp_quarterly['time'] = df_temp_quarterly['time'] + pd.offsets.MonthEnd(0)
df_temp_install = pd.merge(df_temp_quarterly, df_quarterly_solar_install, on='time', how='inner')
df_temp_install_price = pd.merge(df_temp_install, df_quarterly_electricity_price, on='time', how='inner')
df_temp_install_price_gdp = pd.merge(df_temp_install_price, df_gdp_quarterly, on='time', how='inner')
df_temp_install_price_gdp_population = pd.merge(df_temp_install_price_gdp, df_population_quarterly,
on='time', how='inner')
df_temp_install_price_gdp_population_demand = pd.merge(df_temp_install_price_gdp_population,
df_total_demand_quarterly, on='time', how='inner')
```

Exploratory Data Analysis (EDA)

We started to explore the data by plotting the temperature, population, and GDP against the energy demand to see whether there was any trends or correlation between the variables. Following this, we looked at pricing and solar Installations as independent trends over time to see how they were changing. Further EDA and visualisation is available in Appendix 1.

Temperature and Demand

The scatter plot (figure 9) of temperature and demand over time depicts a v-shaped pattern. Below approximately 20 degrees, the demand increases as the temperature decreases; above 20 degrees, the demand rises with increasing temperature. This pattern aligns with our everyday experience. This empirical evidence implies that temperature is suitable explanatory variable for electricity demand.

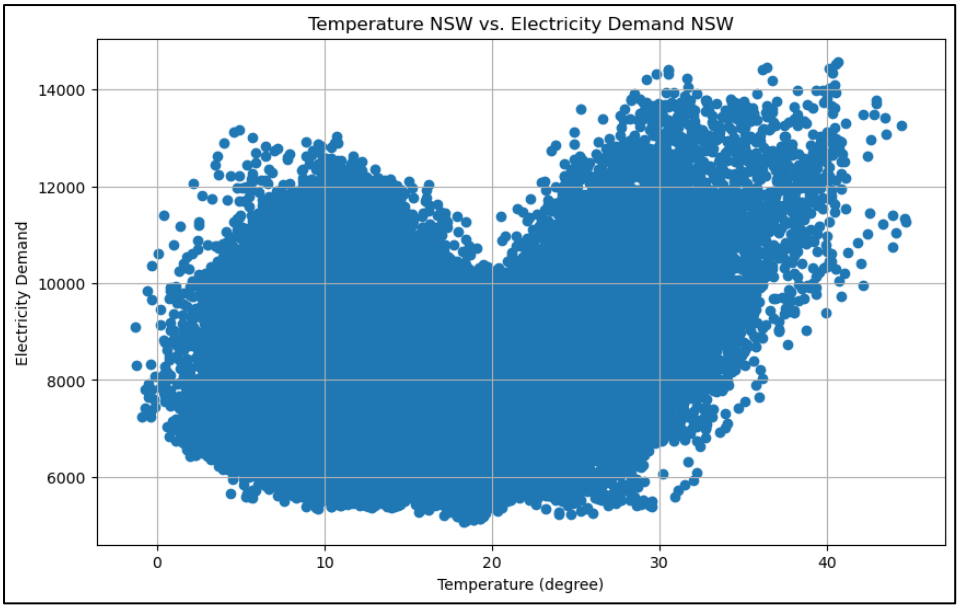


Figure 9: Temperature and Demand

Population and Demand

As the population grows, we would typically expect electricity demand to rise, assuming individual consumption remains steady or increases with technological advances. Contrary to this expectation, we observe (figure 10) a decline in electricity demand with rising population. This suggests that the per capita electricity demand is decreasing. Possible explanations for this trend might include a relative decrease in personal income against the cost of electricity, prompting individuals to limit usage, or the adoption of alternative energy sources, like residential solar panels, reducing reliance on traditional electricity supplies. We were unable to separate residential and industrial demand, so this observation has limited significance.

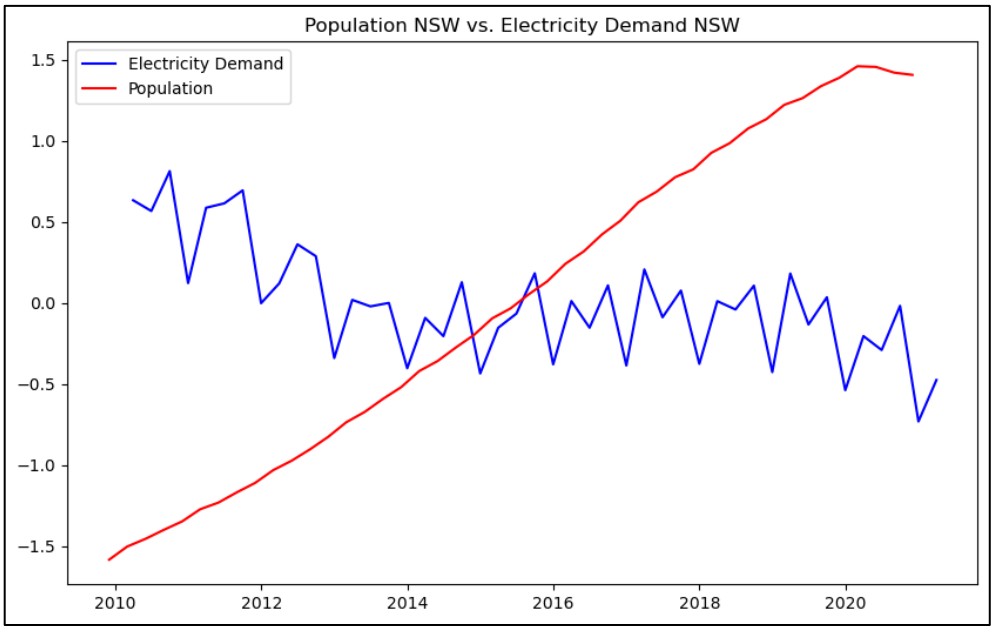


Figure 10: Population and Demand

GDP and Demand

GDP can roughly (if not totally) represent the income of citizens. GDP kept increasing from 2010 to 2020, while demand did not increase, similar to the population trend. From the literature review, this is most likely attributable to the reduced reliance on traditional energy sources, but also an increasing GDP would indicate a population that is increasing in wealth and purchasing more energy efficient appliances, which is backed by our research. We can observe that an increase in electricity demand is followed by an increase in GDP, with a lag, which can be interpreted as increased industrial and economic activities drive higher demand, subsequently followed by higher GDP from such activities.

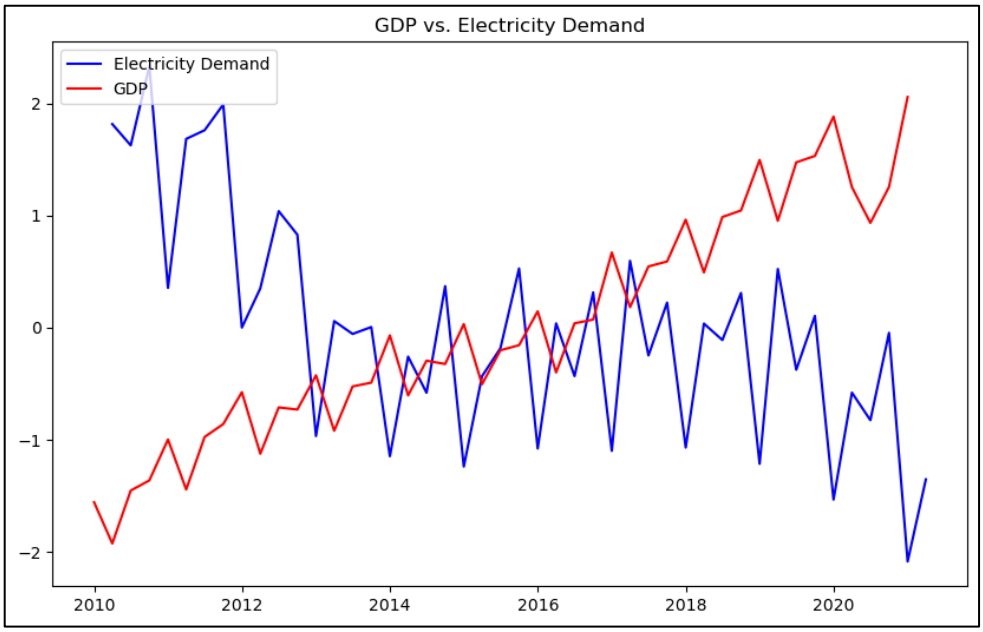


Figure 11: GDP and Demand

Pricing Trend

We used the pricing dataset from CEIC's website. We can see a generally increasing trend in retail price, but volatility is evidently present from year to year.

Figure 12: Pricing over time

Solar PV Installations

We sourced small-scale PV installation postcode data from Clean Energy Regulator, Australian Government. As per the chart below, we can see as expected solar PV installations are trending upwards consistently over the last decade. This supports the observations during literature review that solar PV is accounting for a larger amount of overall energy demand and consumption over time.

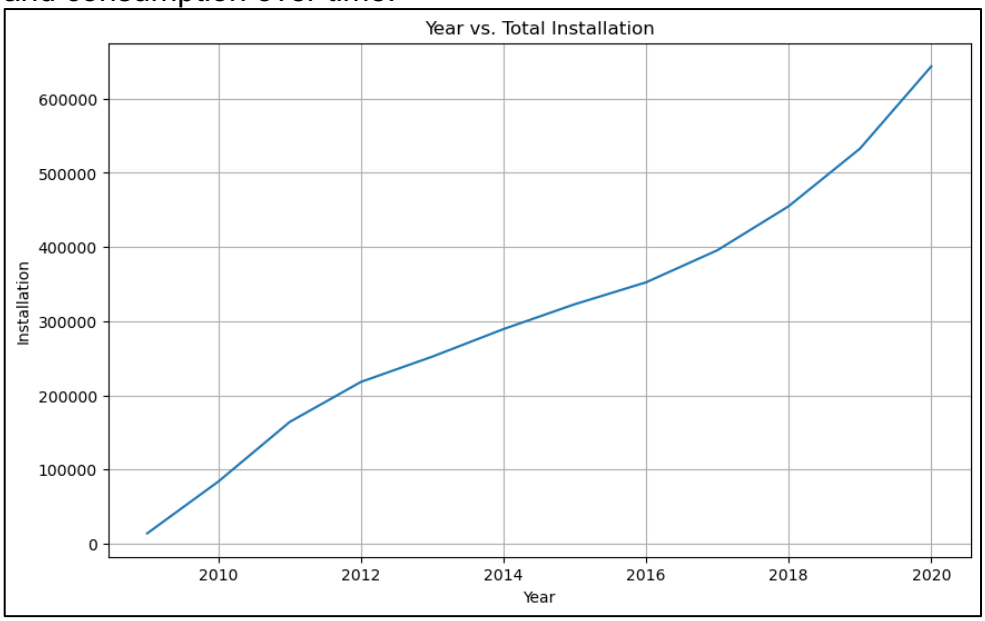


Figure 13: Solar PV Installations in the last decade

Correlation Analysis

To begin we did a correlation analysis to determine which features are best correlated to the total demand.

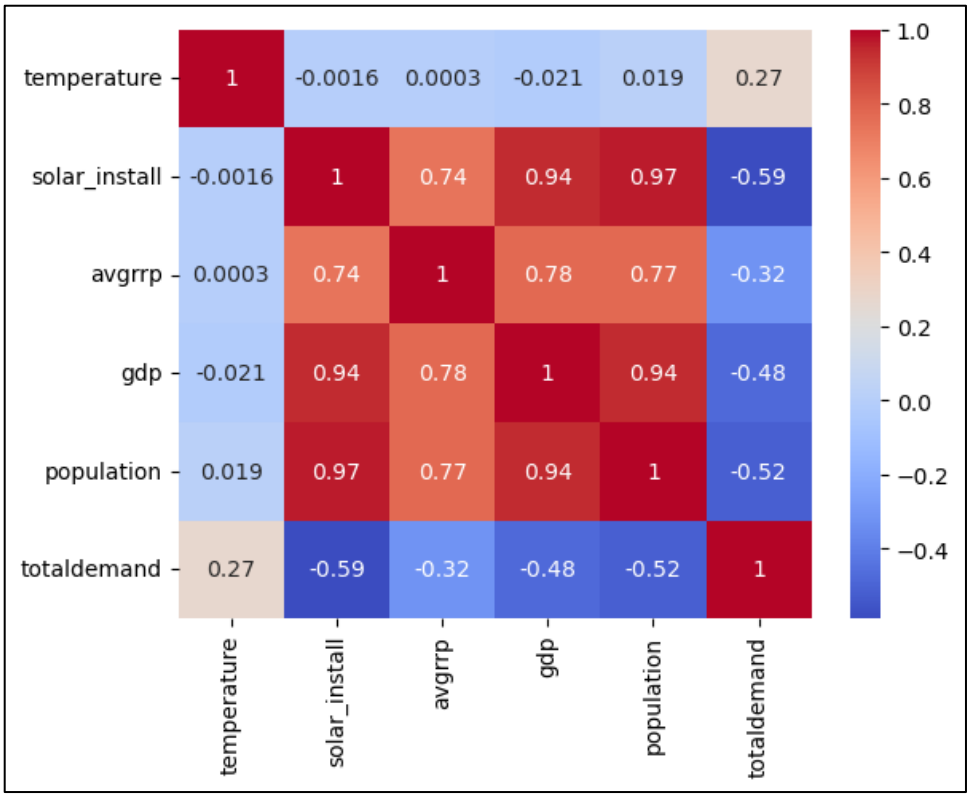


Figure 14: Correlation Analysis

From the heat map (figure 14) we can see that both solar Installations and population are strong influencers of total demand. It is important to distinguish at this point that the total demand represents the demand on the traditional sources of energy and does not account for demand met by alternate sources like personal solar PV installations. Hence population is negatively correlated with total demand. As per the literature review and the Exploratory Data Analysis, this observation is aligned to the downward trend on energy demand against population. To certain extent, this may also be explained by the energy efficiencies that are created through the introduction of renewable energy sources.

Modelling

We looked at two types of energy prediction models:

- **Long Term**: We can predict the average demand for a quarter. As discussed above, such a prediction should consider factors including time, quarterly average temperature, population, GDP, price, and solar installations.
- **Short Term**: We can predict short-term demand every 30 minutes. For such predictions, long-term factors such as population, GDP, price, or solar installations will not play a significant role. Therefore, we will exclude them from our modelling.

Long Term Modelling – Linear Regression

When doing the Linear Regression modelling we excluded the time data. We set demand as the dependent variable, and we had five independent variables:

- Temperature
- Solar Installations

- Price
- GRP
- Population

We have the following function was determined (coefficients rounded to 6 decimal places):

$$y = 30.402235x_1 - 0.003747x_2 + 4.300684x_3 + 0.0008x_4 + 0.00053x_5 + \varepsilon$$

where,

$x_1 = Temperature, -10 \leq x_1 \leq 50$
$x_2 = SolarInstallations, 0 \leq x_1 \leq \infty$
$x_3 = Price, 0 \leq x_1 \leq \infty$
$x_4 = GDP, 0 \leq x_1 \leq \infty$
$x_5 = Population, 0 \leq x_1 \leq \infty$

Evaluation Metrics	Value
Mean Square Error (MSE)	67635.3984478119

Table 9: Evaluation Metrics for the Linear Regression

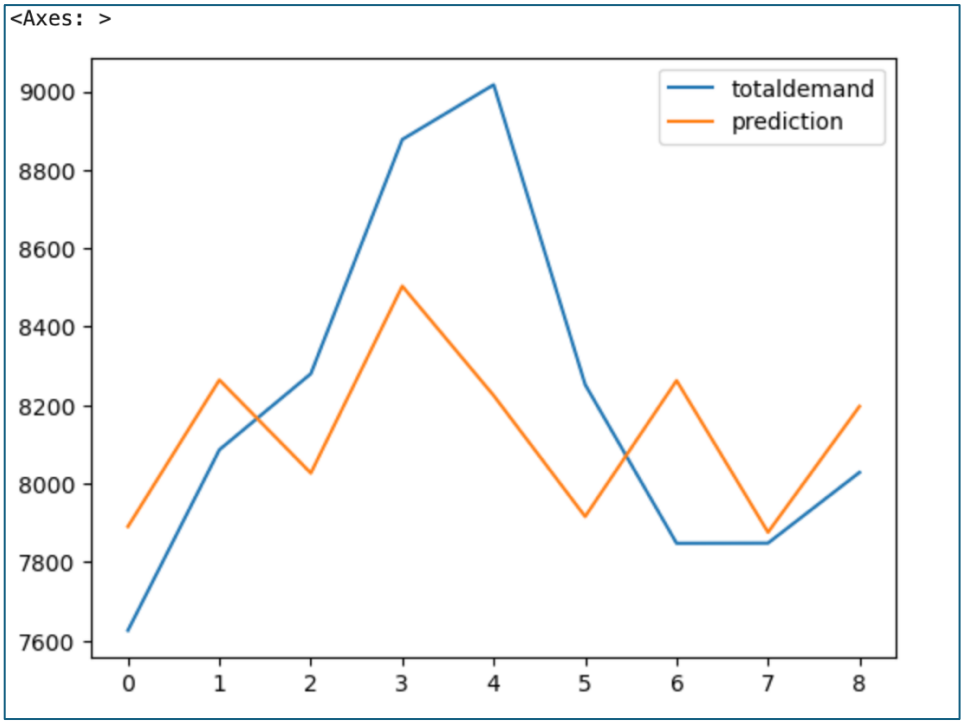


Figure 14: Linear Regression Plot of Total Demand and the Prediction

Long Term Modelling – Polynomial Regression

When predicting quarterly demand, we used population, average temperature, and GDP. We built a polynomial regression model.

Polynomial Regression Modelling

When doing the Polynomial Regression modelling, we excluded the time data. We set demand as the dependent variable, and we had five independent variables:

- Temperature
- Solar Installations
- Price
- GDP
- Population

The table below summarises coefficients of different predictors along with p-values of each. The predictors that were found to be statistically significant at 95% confidence level are identified with an asterisk (*) next to their p-value. For example, second order term of temperature is the most significant term, being significant at 99% confidence level. Whereas population is not found to be significant in this model. As demand tends to increase at temperatures below or higher than certain thresholds (i.e. outside 'normal' temperature), it can be assumed that the relationship between temperature and electricity demand will be non-linear and most likely quadratic.

	Coefficient	Standard Error	t-Statistic	p-Value
Intercept	0.20	0.00	0.00	0.999999
temperature	25.50	0.08	0.03	0.974475
solar_install	-0.21	0.17	-1.28	0.227917
avgrrp	-315.00	298	-1.06	0.313624
gdp	0.53	0.33	1.56	0.146258
population	-0.02	0.08	-0.26	0.802009
temperature^2	-42.20	5.45	-7.75	0.000009*
temperature solar_install	-0.00	0.00	-2.62	0.024038*
temperature avgrrp	-1.82	0.06	-2.92	0.014034*
temperature gdp	0.00	0.00	3.34	0.006597*
temperature population	0,00	0.00	0.36	0.72177
solar_install^2	0.00	0.00	-2.14	0.055699
solar_install avgrrp	0.00	0.00	0.01	0.959917
solar_install gdp	0.00	0.00	2.41	0.03473*
solar_install population	0.00	2.23E-08	0.76	0.460923
avgrrp^2	-0.38	0.10	-3.91	0.002448*
avgrrp gdp	0.00	0.00	-0.23	0.823291
avgrrp population	0.00	0.00	1.24	0.239283
gdp^2	0.00	0.00	-0.48	0.641729
gdp population	0.00	0.00	-1.58	0.142371
population^2	0.00	0.00	0.55	0.5932

Table 10: Results of Polynomial Regression rounded to 2 decimal places

The following evaluation metrics were identified for the polynomial regression. It is notable that the training dataset produced high R^2 but the testing data did not. Possibility of overfitting may be explored further to understand the testing performance.

Evaluation Metrics	Value (rounded to 2 decimal places)
Training Mean Square Error (MSE)	12,755.87
Testing Mean Square Error (MSE)	80,677.61
Training R - Squared	0.91
Testing R - Squared	0.56

Table 11: Evaluation Metrics for the Polynomial Regression

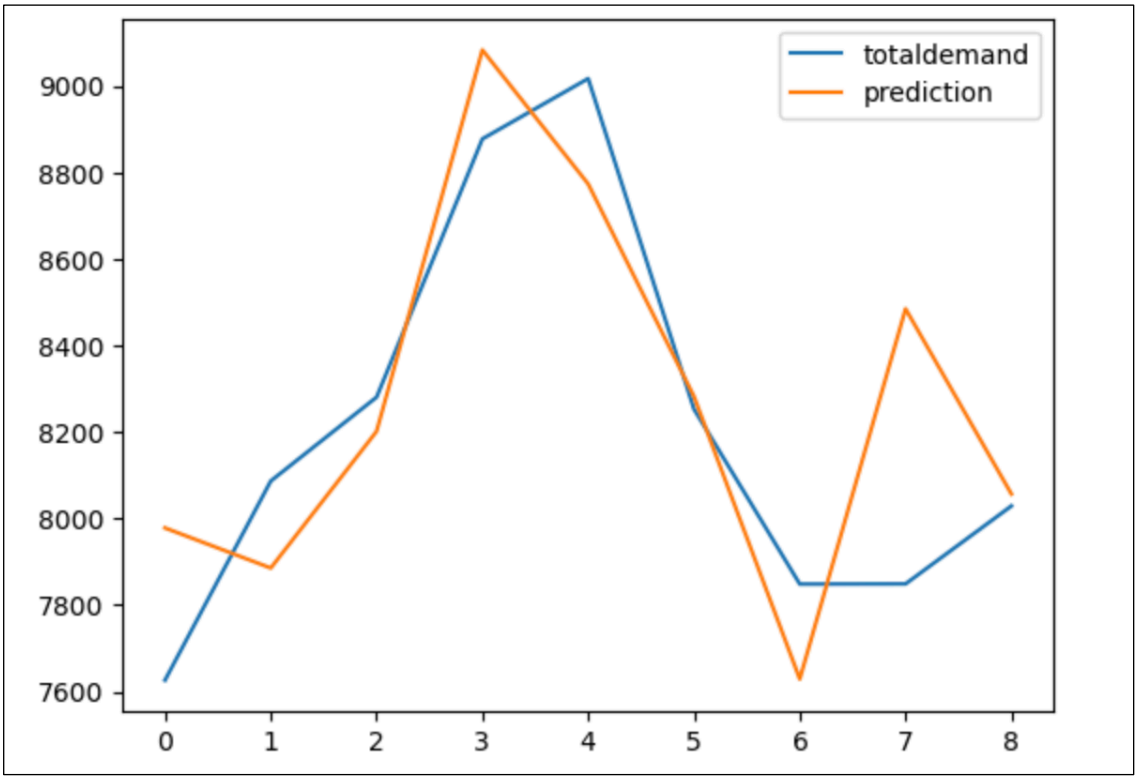


Figure 15: Polynomial Regression Plot of Total Demand and the Prediction

Short Term Modelling – ARIMA

For short-term modelling we investigated the impact of temperature on the energy demand and use historical demand and temperature data over a time series to predict the future energy demand. We first approached this by using the full dataset, to see if by having a large data set that we could get an accurate prediction, before tuning the data further.

First, the data was evaluated for any presence of stationarity. While in absence of stationarity an ARMA[2] (p,q) or ARIMA (p.0,q) model may be appropriate, presence of stationarity pose further complexity.

[2] Autoregressive Moving Average

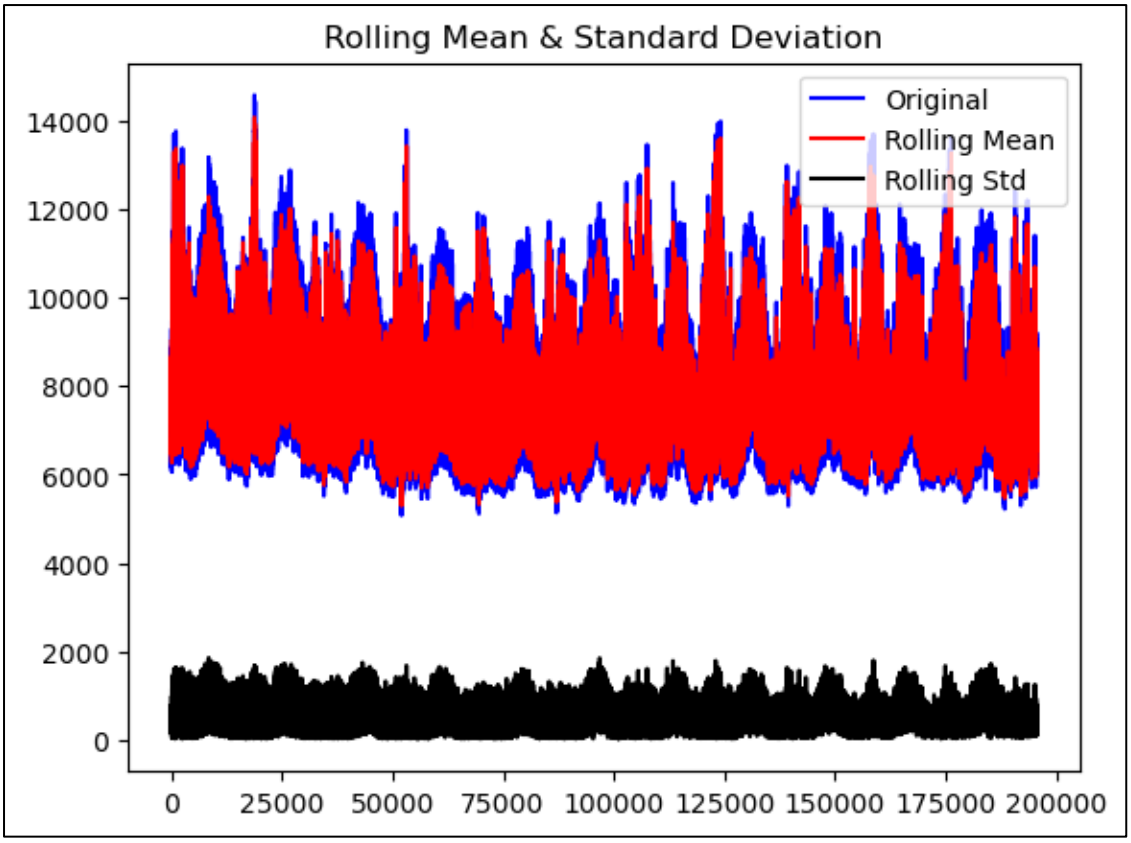


Figure 16: Rolling Mean and Standard Deviation

Dickey-Fuller test resulted in the following evaluation metrics:

- Test Statistic: -25.48
- p-value: 0.000000
- #Lags Used: 80
- Number of Observations Used: 195,530
- Critical Value (1%): -3.43
- Critical Value (5%): -2.86
- Critical Value (10%): -2.57

The low p-value from the test and test statistic being far below the critical range imply that the null hypothesis, that the data is stationary, can be rejected at 99% confidence level. This is explored in more details under Discussion section. The non-stationarity issue can normally be addressed by introducing a differencing term in the ARIMA, the extent of differencing being such that it eliminates any non-stationarity in the time series. However, due to computational constraints this was not introduced in the model and therefore, the impact of stationarity in performance on the model is evident in the predictions.

The next step in preparing an ARIMA model is to determine the appropriate parameters:

- p: autoregressive term
- d: differencing order
- q: moving average term

As mentioned above, despite strong indication of presence of stationarity in the data, we set d=0 for rest of the modelling to tackle computational limitations of the ARIMA model. This in vestigation notes that the performance of this model may significantly improve with introduction of a differencing term of appropriate order. We ran a grid search to determine the optimal p and q values, and arrived at p = 3 and q = 4 which had th e lowest MAPE[3] (noted in table 4 below).

Evaluation Metric	Value
MAE (Mean Absolute Error)	744.89
MAPE (Mean Absolute Percentage Error)	0.0957
MSE (Mean Square Error)	827,112.78

Table 12: Evaluation Metrics for p = 3 and q = 4 on full dataset

The MAPE provides the error as a percentage of the actual value. A lowest MAPE therefore provides the lowest deviation as a percentage of the prediction from the actual. The MAPE is a useful metric due to its simplicity and ease of understanding the accuracy of predictions. However MAPE accuracy can be impacted by extreme values, and hence can make the least MAPE may not always indicate a good fit. This is reflected in figure 16 below, which shows the model is not strong in predicting demand volatility.

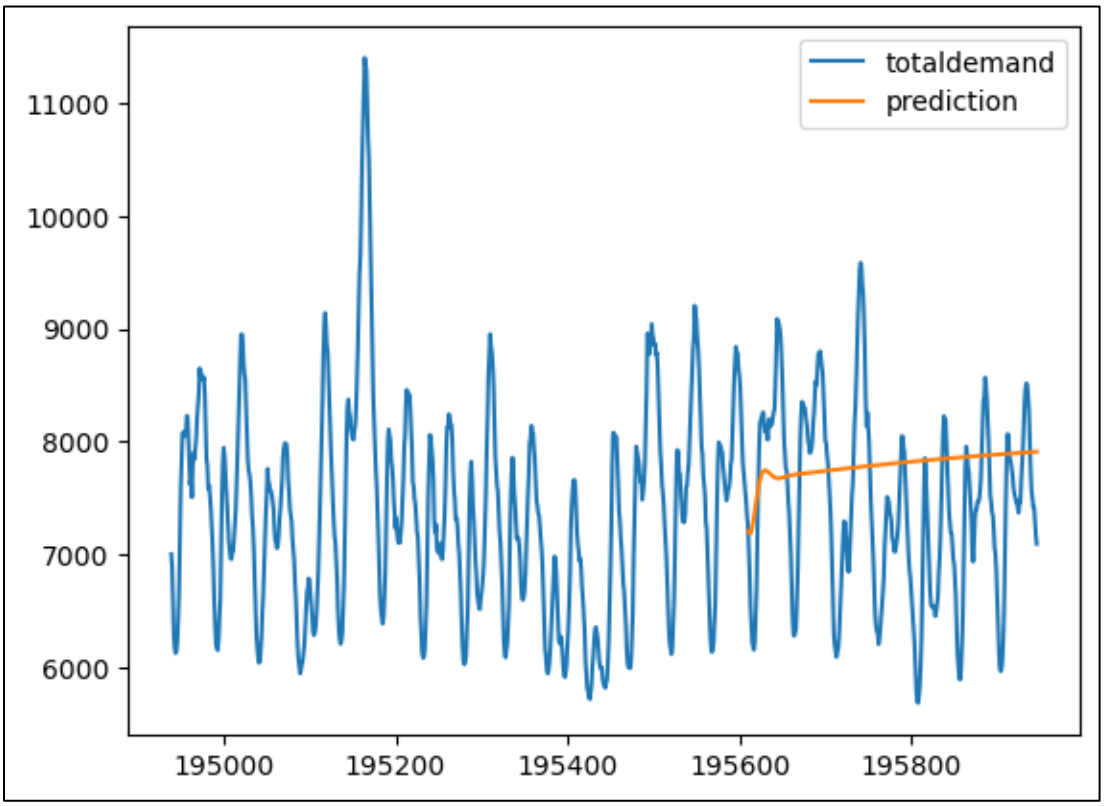


Figure 17: Plot of ARIMA prediction using full dataset

Considering that the full data would be affected by any underlying trend, and to circumvent the limitation of not introducing a differencing factor, the dataset is subsequently limited to a

[3] Mean Absolute Percentage Error

shorter period - 3 months prior to the prediction timeline. Running the ARIMA model again, and performing a grid search, we were able to find an improved prediction with p = 4 and q = 24 (d = 0).

Evaluation Metric	Value
MAE (Mean Absolute Error)	607.58
MAPE (Mean Absolute Percentage Error)	0.0804
MSE (Mean Square Error)	577,072.42

Table 13: Evaluation Metrics for p = 4 and q = 24 on 3 months data

The MAPE in this instance was smaller than previous model. From figure 17, we can see the prediction is a better fit than when using the full dataset. However, it should be noted that the limitations of ARIMA model and parameter selection still applies. Limiting the observation to a significantly shorter period may have removed some trend in the data, leading to better performance of the model. This improved performance also indicates that further parameter tuning, particularly introducing adequate differencing term, can be expected to provide marked improvement in accuracy of prediction. In addition, a known limitation of ARIMA model is failure to predict at extrema where sharp changes precede the maxima. Temperature, and subsequently electricity demand have peaks and troughs followed by sharp changes; therefore, a well-tuned ARIMA may fail to predict the extrema while performing very well otherwise.

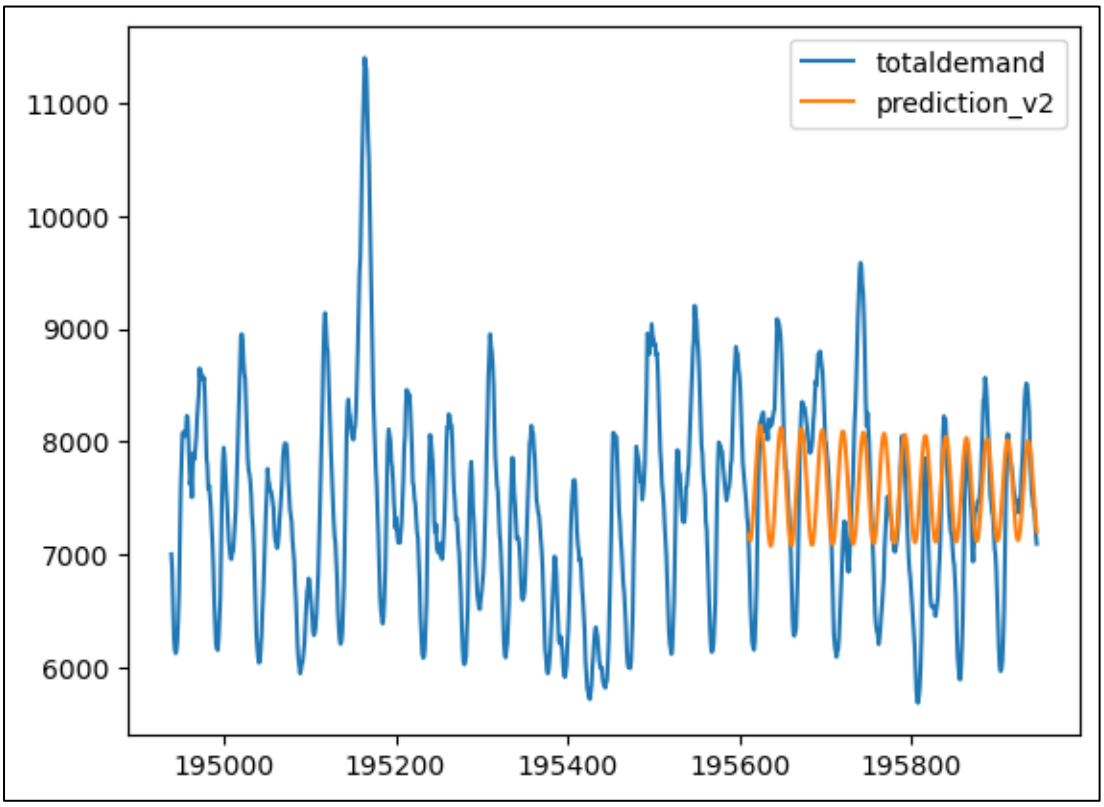


Figure 18: Plot of ARIMA prediction using full dataset

Discussion

Long Term Modelling – Regression Analysis

We started with linear regression modelling because it is easy to understand and interpret, and it works well with small datasets. We see from the plot in Figure 14 a prediction model that provides and indicative but not accurate prediction of demand. This is useable in the long term, as we are not looking to predict an exact number, but rather an indicative range that would enable users to make policy decisions. Having said that, the MSE is still higher than we would like. This may be due to the non-linearity of temperature. Since, the relationship between temperature and demand is clearly V-shaped, a polynomial regression with a degree of 2 is likely more suitable here.

We then looked at a second-degree polynomial regression which may be a closer fit. But we can see the prediction is still not a strong goodness of fit, but from the plot we can see the fit is enough that the model is informative. The testing R-squared error and testing MSE also indicate that the goodness of fit is not strong. We would want to see a testing MSE closer to 0, and a testing R-Squared error closer to 1. Like the linear regression, we are not getting the level of accuracy in the model that we would like, but believe the model is still usable for long term predictions as the error would naturally be much larger. The polynomial regression has a better goodness of fit compared to then linear regression.

Short Term Modelling - ARIMA

Our prediction using ARIMA is useful but limited. It provides a usable prediction but has scope of further parameter refinement to improve predictions, aside general limitation of ARIMA models at the extremes. With time and computational capability, we could address stationarity and refine parameters achieve substantial accuracy in prediction.

While we have confidence on the capability of ARIMA model in electricity demand prediction to a great extent, for energy suppliers who may need very high level of accuracy in short time intervals, ARIMA model may not be ideal. This is backed by the literature review, and in particular the study Finding the Best ARIMA Model to Forecast Daily Peak Electricity Demand (As'ad 2012) which explains that the ARIMA Model is not effective for data with extremes. For energy demand this is an issue as predicting the extremes is most useful, as that is when there will be the most demand for energy.

By producing an Auto-Correlation Function (ACF) and Partial Auto-Correlation Function (PACF) plots to assess the seasonality of the data.

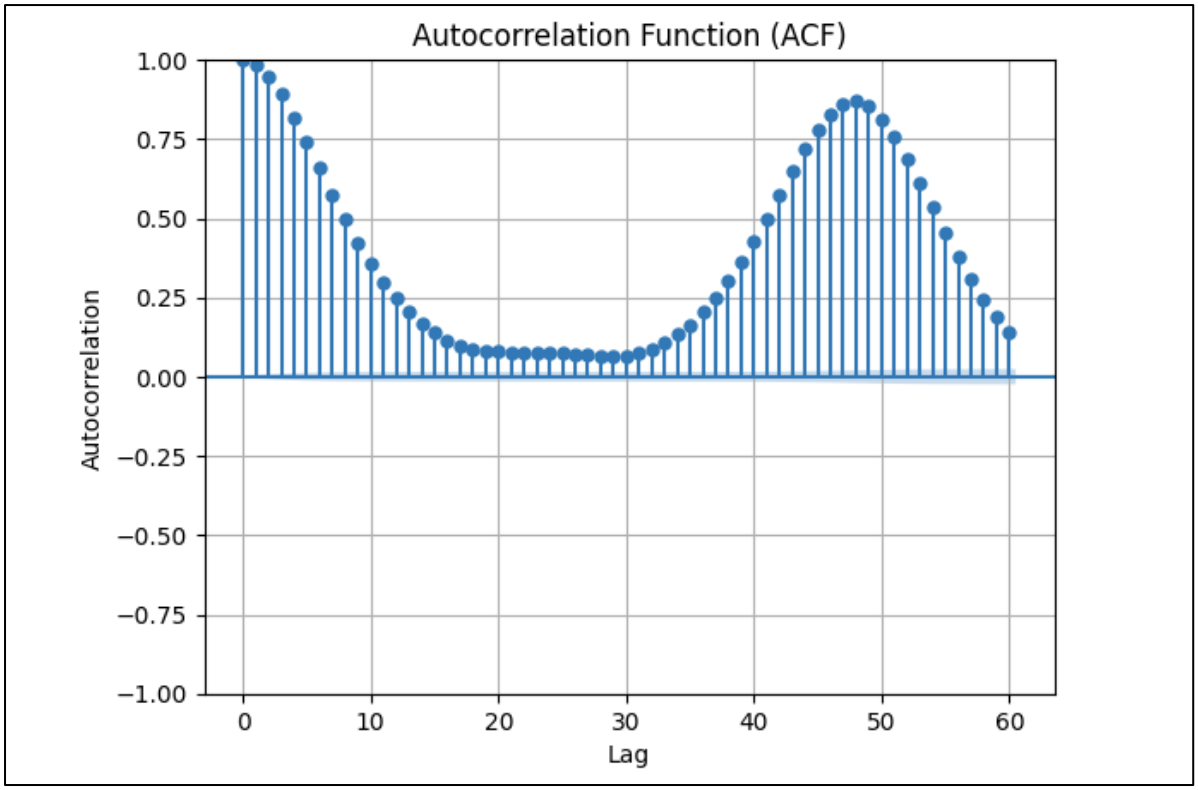


Figure 18: ACF Plot

The ACF plot indicates three key findings:

- Strong Seasonality - The ACF plot exhibits a pronounced cyclical pattern characterized by peaks and troughs at consistent intervals. This pattern is indicative of a significant seasonal influence within the temperature data. Such a recurring pattern over intervals suggests the need for a seasonal adjustment in our time series analysis.
- Seasonal Cycle Identification - The interval between successive peaks in the ACF plot represents the length of the seasonal cycle. Identifying this length is critical as it informs the seasonal component of a SARIMA (Seasonal ARIMA) model. The regularity and significance of these peaks suggest the presence of a strong seasonal component that should be accounted for in the modelling process.
- Gradual Decline in Autocorrelation - A slow decrease in the autocorrelation values as the lag increases suggests potential non-stationarity in the series. Non-stationarity implies that the statistical properties of the series change over time, which is a crucial consideration for time series modelling as most models assume stationarity.

The ACF plot has provided valuable insights that direct the construction of an effective time series model. The strong seasonality and potential non-stationarity imply the use of SARIMA modelling with adequate differencing to stabilize the mean. The identification of the seasonal cycle will inform the seasonal differencing term to ensure that the model captures the underlying patterns accurately.

Similarly, by looking at the PACF plot in figure 19 we get four key findings:

- Immediate Lag Significance - The significant spike at the immediate first lag indicates a strong correlation that is not explained by the correlation at all other lags.

This suggests that the temperature data has a significant autoregressive component at lag 1.

- Tail-off Pattern - The PACF plot shows a tailing off pattern, with most lags beyond the first not showing significant partial autocorrelation. This tail-off indicates that the direct correlation between the temperature data points and their subsequent lags diminishes quickly, which is typical for an AR(1) process where only the first lag is considered.
- Lag Order for AR Model - The significance of the initial lag in the PACF plot is a strong indication of the need for an AR(1) term in the ARIMA model. It suggests that the immediate past value has a substantial impact on the current value of the series, with minimal influence from further past values.
- Negative Correlation at Higher Lags - The PACF plot does not show any significant negative correlations at higher lags that would be indicative of an oscillating pattern which is typically seen in over-differenced series.

The PACF analysis confirms that an AR(1) model is appropriate for the temperature time series data. The significant correlation at lag 1 and the rapid decrease in correlation at higher lags provide clear guidance for the AR component of our predictive model. The lack of significant correlations at higher lags suggests that the temperature data is likely not over-differenced.

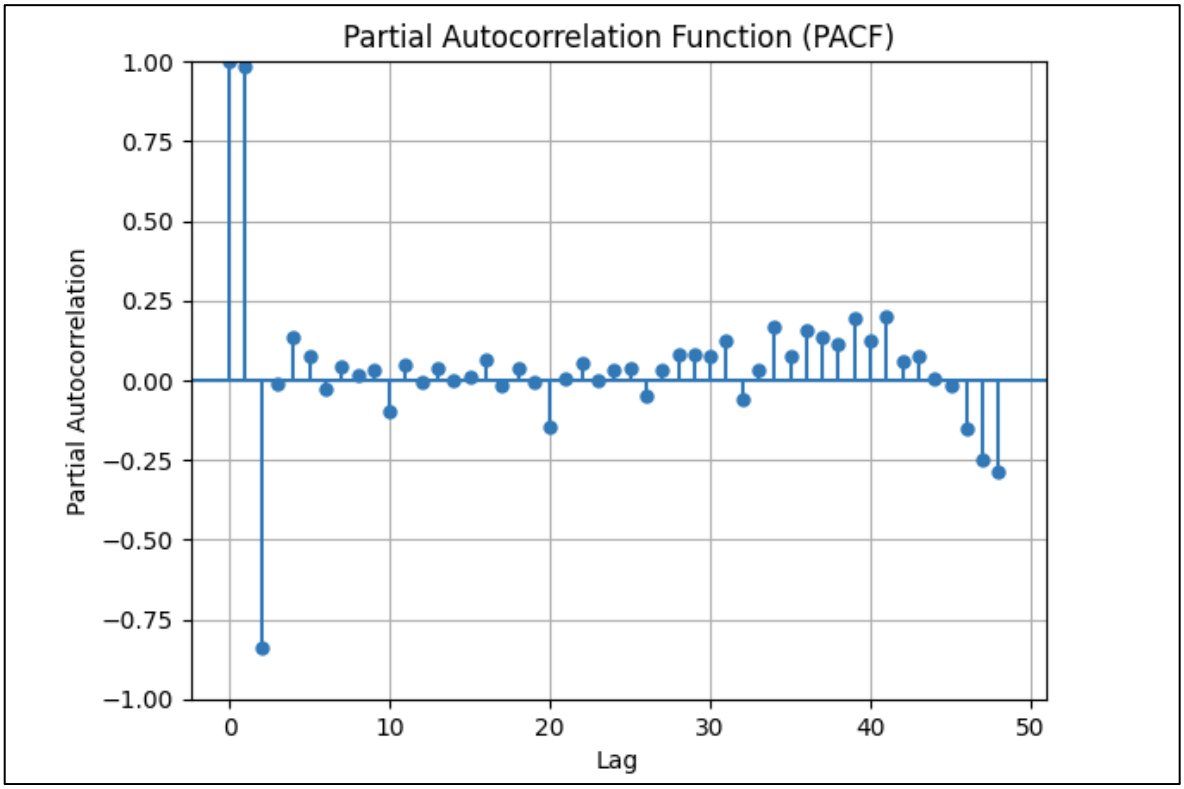


Figure 19: PACF plot

Conclusion and Further Issues

The results in this paper are informative to understand influencing factors for both long-term and short-term, which can support forecasting and decision making.

From a long-term perspective, three key factors which this paper determined to strongly influence energy demand were temperature, solar PV installations and prics.

First, policy makers can use the linear regression model to understand at a basic level the impact Solar PV Installations has on the reduction in energy demand from traditional sources (Coal and Natural Gas). By doing scenario-based planning to determine the optimal amount of solar PVs that will be required to meet our future energy needs, the NSW Government can develop policies and incentive schemes targeted at increasing grass-roots solar PV installations to achieve the renewable energy targets. This analysis, in particular the findings from the Literature Review in regard to the energy efficiencies generated through solar PVs and other renewable sources, and Exploratory Data Analysis, will be useful for driving support for climate change policy discussions and analysis more holistically.

Second, the NSW Government can use the regression model to understand the impact population has on the energy demand. By using their own population projection, they will be able to see what the long-term energy demand is going to look like, also taking into consideration GDP, and price as other factors that would drive changes to the energy demand along with the population.

From a short-term perspective, temperature is the key element in influencing energy demand. Energy distributors and organisations in the electricity industry could use this model to predict energy demand. By doing so, they can incentivise their customers through a variety of means, including but not limited to, schemes to reduce load during hot hours of the day, or pre-purchasing energy blocks so smooth out demand over a period time. The ARIMA model shows that by using a shorter time period (3 months) of data, we can predict with confidence a short span forward. The limitation around the accuracy of the ARIMA models for extreme values, means while the model is informative, and can be used to drive customer behaviour through incentives, there are likely more useful models that better predict the demand at temperature extremes.

Following are few areas that we would have liked to explore further, if time permitted.

1. The impact of solar PVs and investment in energy storage facilities on the overall energy demand to substantiate current renewable energy policy or suggest improvement. By looking at data that distinguished residential solar PV installations versus industrial scale solar PV installations, and how more large-scale implementations of solar PVs could impact energy demand would be more useful for policy makers and investors to make informed decisions.

2. More investigation on impact of retail pricing. Given the macroeconomic environment in 2023 and 2024, and prevailing high inflation rate in Australia, there has been a lot of discussion about affordability of basic bills (i.e. rent, utilities, etc.). Our model showed that price does influence energy demand over the long term; but given more time, we should investigate this more detail with other models to ensure price affordability for all NSW residents.

3. Although not included in the main body of this report, we undertook a modelling exercise using XGBoost (XGB) and Long Short-Term Memory (LSTM) networks to explore additional areas of study. Unfortunately, due to a lack of comprehensive data, this exercise did not yield insights suitable for inclusion in the final report. Nevertheless, the methodologies and processes involved in the modelling have been documented in the appendix section of this report. Additionally, the corresponding code has been made available in our GitHub repository.

4. A more robust ARIMA model, that addresses non-stationarity in the data by introducing a differencing term (based on auto-correlation factor analysis) along with further tuning of parameters to achieve better fit (higher R^2 and lower residual variance at testing phase).

References

Ahmed, T., Muttaqi, K.M. and Agalgaonkar, A.P. (2012). Climate change impacts on electricity demand in the State of New South Wales, Australia. *Applied Energy*, 98, pp.376–383. doi:https://doi.org/10.1016/j.apenergy.2012.03.059.

Akhwanzada, S.A. and Tahar, R.M. (2012). Strategic Forecasting of Electricity Demand Using System Dynamics Approach. *International Journal of Environmental Science and Development*, [online] 3(4). Available at: https://www.ijesd.org/papers/241-B063.pdf [Accessed 15 Apr. 2024].

As'ad, M. (2012). Finding the Best ARIMA Model to Forecast Daily Peak Electricity Demand. *Applied Statistics Education and Research Collaboration (ASEARC) - Conference Papers*. [online] Available at: https://ro.uow.edu.au/asearc/12 [Accessed 15 Apr. 2024].

Australian Energy Market Operator (2023a). *Major Publications Library*. [online] aemo.com.au. Available at: https://aemo.com.au/-/media/files/major-publications/isp/2023/2023-iasr-infographic.pdf?la=en [Accessed 15 Apr. 2024].

Australian Energy Market Operator (2023b). *NEM Electricity Statement of Opportunities ESOO*. [online] aemo.com.au. Available at: https://www.aemo.com.au/-/media/files/electricity/nem/planning_and_forecasting/nem_esoo/2023/2023-electricity-statement-of-opportunities.pdf?la=en [Accessed 15 Apr. 2024].

Department of Climate Change, Energy, the Environment and Water (2023a). *National Energy Performance Strategy - DCCEEW*. [online] Dcceew.gov.au. Available at: https://www.dcceew.gov.au/energy/strategies-and-frameworks/national-energy-performance-strategy [Accessed 15 Apr. 2024].

Department of Climate Change, Energy, the Environment and Water (2023b). *Powering Australia - DCCEEW*. [online] Dcceew.gov.au. Available at: https://www.dcceew.gov.au/energy/strategies-and-frameworks/powering-australia [Accessed 15 Apr. 2024].

Energy Security Board (2021). *Integration of consumer energy resources (CER) and flexible demand*. [online] ESB. Available at: https://esb-post2025-market-design.aemc.gov.au/integration-of-distributed-energy-resources-der-and-flexible-demand [Accessed 18 Mar. 2024].

Energy.gov.au. (2022). *National Energy Transformation Partnership | energy.gov.au*. [online] Available at: https://www.energy.gov.au/energy-and-climate-change-ministerial-council/national-energy-transformation-partnership [Accessed 15 Apr. 2024].

Fong, W.K., Matsumoto, H., Lun, Y.F. and Kimura, R. (2007). System dynamic model for the prediction of urban energy consumption trends. Proceedings I of the 6th international conference on indoor air quality. *ventilation & energy conservation in buildings*, pp.762–769.

Ghalehkhondabi, I., Ardjmand, E., Weckman, G.R. and Young, W.A. (2016). An overview of energy demand forecasting methods published in 2005–2015. *Energy Systems*, [online] 8(2), pp.411–447. doi:https://doi.org/10.1007/s12667-016-0203-y.

Hagan, M.T. and Behr, S.M. (1987). The Time Series Approach to Short Term Load Forecasting. *IEEE Transactions on Power Systems*, 2(3), pp.785–791. doi:https://doi.org/10.1109/tpwrs.1987.4335210.

Hor, C.-L., Watson, S.J. and Majithia, S. (2006). Daily Load Forecasting and Maximum Demand Estimation using ARIMA and GARCH. *2006 International Conference on Probabilistic Methods Applied to Power Systems*. [online] doi:https://doi.org/10.1109/pmaps.2006.360237.

IEA (2019). *Data & Statistics - IEA*. [online] International Energy Agency. Available at: https://www.iea.org/data-and-statistics [Accessed 15 Apr. 2024].

Intarapravich, D., Johnson, C.J., Li, B., Long, S., Pezeshki, S., Prawiraatmadja, W., Tang, F.C. and Wu, K. (1996). 3. Asia-Pacific energy supply and demand to 2010. *Energy*, [online] 21(11), pp.1017–1039. doi:https://doi.org/10.1016/0360-5442(96)00085-0.

International Energy Agency. (n.d.). *International Energy Agency*. [online] Available at: https://www.iea.org/about [Accessed 15 Apr. 2024].

Islam, M.A., Che, H.S., Hasanuzzaman, M. and Rahim, N.A. (2020). Energy demand forecasting. *Energy for Sustainable Development*, pp.105–123. doi:https://doi.org/10.1016/b978-0-12-814645-3.00005-5.

Kareem, Y.H. and Majeed, A.R. (2006). Monthly Peak-load Demand Forecasting for Sulaimany Governorate Using SARIMA. doi:https://doi.org/10.1109/tdcla.2006.311383.

Kayacan, B., Ucal, M., Öztürk, A., Balı, R., Koçer, S. and Kaplan, E. (2012). A primary econometric approach to modelling and forecasting the demand for fuelwood in Turkey. *Journal of Food Agriculture and Environment*, 1010, pp.934–937.

Latief, Y. (2023). *Power outage potential drives updated Australian market rules*. [online] Smart Energy International. Available at: https://www.smart-energy.com/industry-sectors/business/australia-updates-market-rules-to-reduce-power-outage-potential/ [Accessed 18 Mar. 2024].

McLoughlin, F., Duffy, A. and Conlon, M. (2013). Evaluation of time series techniques to characterise domestic electricity demand. *Energy*, 50, pp.120–130. doi:https://doi.org/10.1016/j.energy.2012.11.048.

Mohamed, Z. and Bodger, P. (2005). Forecasting electricity consumption in New Zealand using economic and demographic variables. *Energy*, 30(10), pp.1833–1843. doi:https://doi.org/10.1016/j.energy.2004.08.012.

Nelson, T. (2015). Australian Climate Change Policy - Where To From Here? *Economic Papers: A journal of applied economics and policy*, 34(4), pp.257–272. doi:https://doi.org/10.1111/1759-3441.12114.

Nogales, F.J., Contreras, J., Conejo, A.J. and Espinola, R. (2002). Forecasting next-day electricity prices by time series models. *IEEE Transactions on Power Systems*, 17(2), pp.342–348. doi:https://doi.org/10.1109/tpwrs.2002.1007902.

NSW Climate and Energy Action (2022). *Electric Vehicle Strategy*. [online] Available at: https://www.energy.nsw.gov.au/sites/default/files/2022-09/nsw-electric-vehicle-strategy-210225.pdf [Accessed 15 Apr. 2024].

NSW Environment Protection Authority (n.d.). *Energy Consumption | NSW State of the Environment*. [online] www.soe.epa.nsw.gov.au. Available at: https://www.soe.epa.nsw.gov.au/all-themes/human-settlement/energy-consumption [Accessed 15 Apr. 2024].

Sandiford, M., Forcey, T., Pears, A. and McConnell, D. (2015). Five Years of Declining Annual Consumption of Grid-Supplied Electricity in Eastern Australia: Causes and Consequences. *The Electricity Journal*, 28(7), pp.96–117. doi:https://doi.org/10.1016/j.tej.2015.07.007.

Sigauke, C. and Chikobvu, D. (2011). Prediction of daily peak electricity demand in South Africa using volatility forecasting models. *Energy Economics*, 33(5), pp.882–888. doi:https://doi.org/10.1016/j.eneco.2011.02.013.

Smith, M., Hargroves, K. (Charlie), Stasinopoulos, P., Stephens, R., Desha, C. and Hargroves, S. (2007). *Energy transformed: Sustainable energy solutions for climate change mitigation*. [online] *eprints.qut.edu.au*. Australia: The Natural Edge Project, CSIRO, and Griffith University. Available at: https://eprints.qut.edu.au/85180/.

Vaudreuil, M.P.: System dynamics computer simulation modelling to forecast the energy demands for the Montachusett region under a Variety of Simulations and Scenarios (Doctoral dissertation, WORCESTER POLYTECHNIC INSTITUTE) (2011)

Verdejo, H., Awerkin, A., Becker, C. and Olguin, G. (2017). Statistic linear parametric techniques for residential electric energy demand forecasting. A review and an implementation to Chile. *Renewable and Sustainable Energy Reviews*, [online] 74, pp.512–521. doi:https://doi.org/10.1016/j.rser.2017.01.110.

Zhang, Q., Ou, X., Yan, X. and Zhang, X. (2017). Electric Vehicle Market Penetration and Impacts on Energy Consumption and CO2 Emission in the Future: Beijing Case. *Energies*, 10(2), p.228. doi:https://doi.org/10.3390/en10020228.

Appendix 1 – ML Modelling

Objective

The objective of this project is to forecast both total grid demand and energy generated from solar panels in New South Wales (NSW). Ideally, we would like to find out when energy generated from solar panels will intersect with total grid demand.
We set out with the goal of

Datasets

In total, 4 datasets were used:

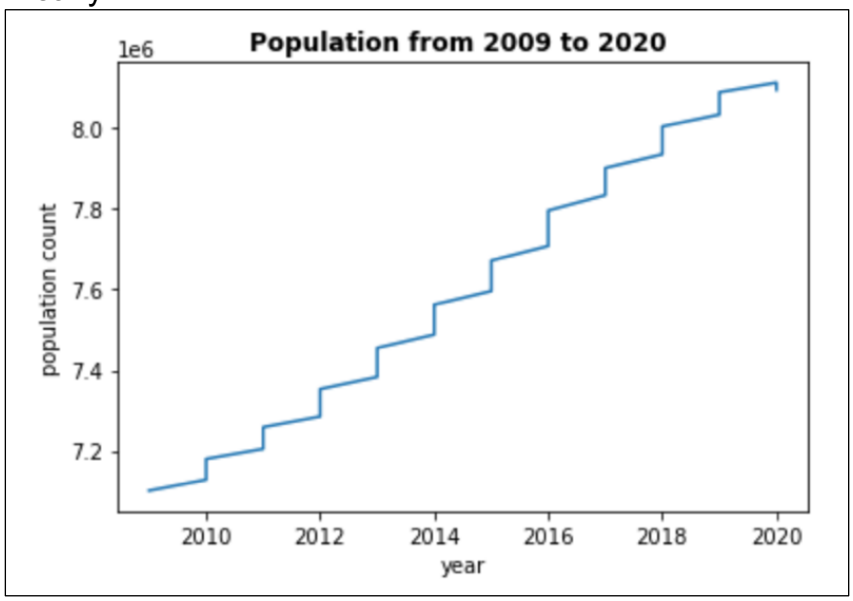
1. population from 2009 to 2020 (quarterly),
2. projected population from 2022 to 2071 (yearly)
3. total grid demand from 2010 to 2021 (every 30 minutes)
4. energy generated from solar panels from 2009 to 2024 (yearly)

Data Cleaning

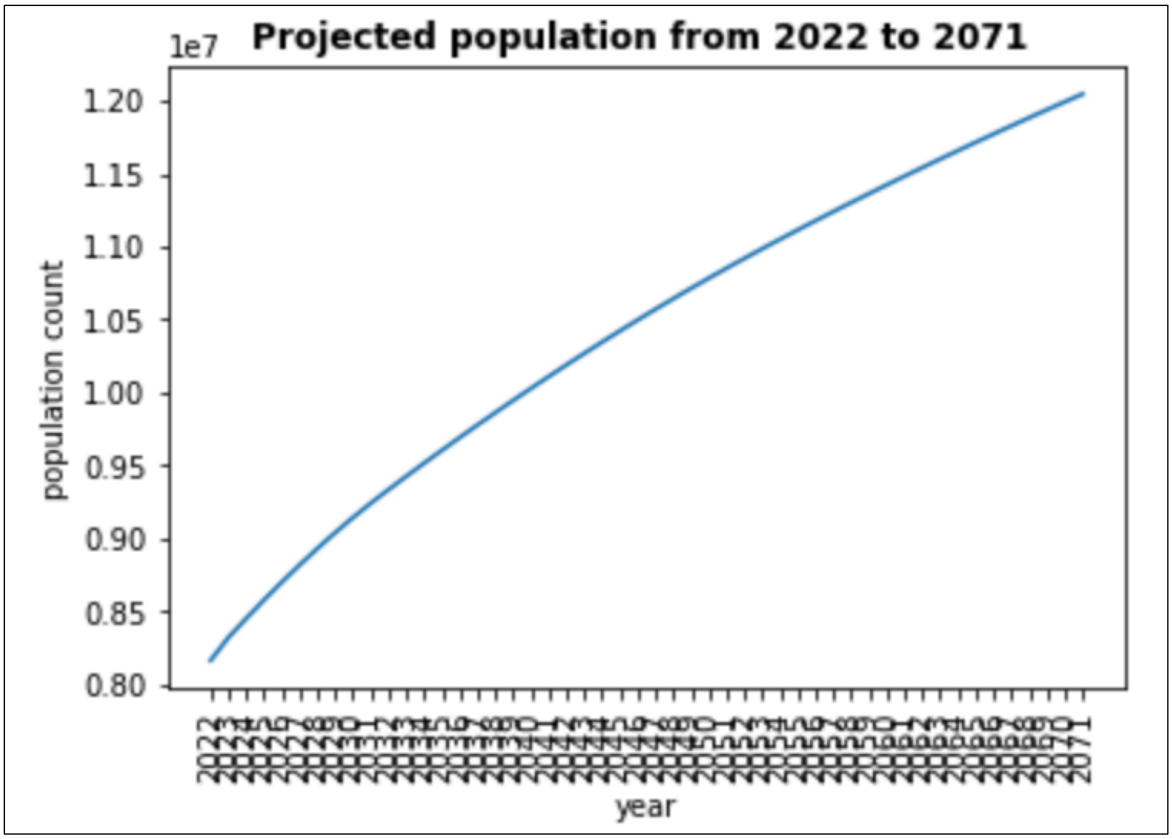
Firstly, datetime columns were converted to datetime format. Also, datetime features such as month, day, hour and minute were created for relevant datasets for further processing.

Data Exploration

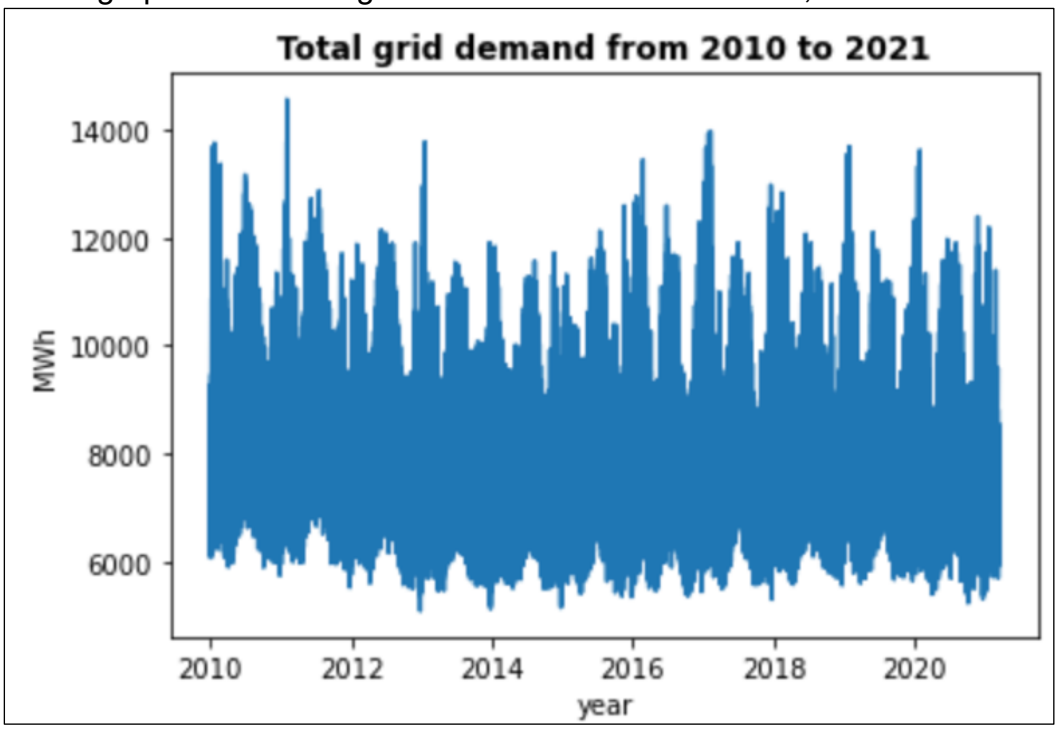
Graph below shows population from 2009 to 2020, where the general trend is increasing linearly.



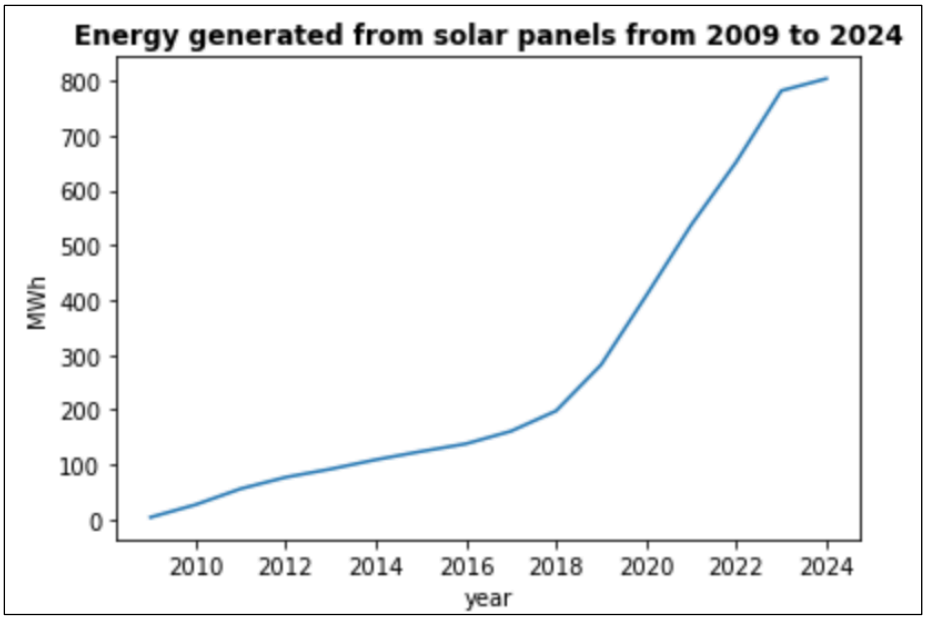
Below graph shows population from 2009 to 2020, where the general trend is increasing logarithmically.



Below graph shows total grid demand from 2010 to 2021, where seasonality is observed.



Below graph shows energy generated from solar panels from 2009 to 2024, where the general trend is increasing exponentially. It is to note that energy generated from solar panels was more than ten times smaller than total grid demand.



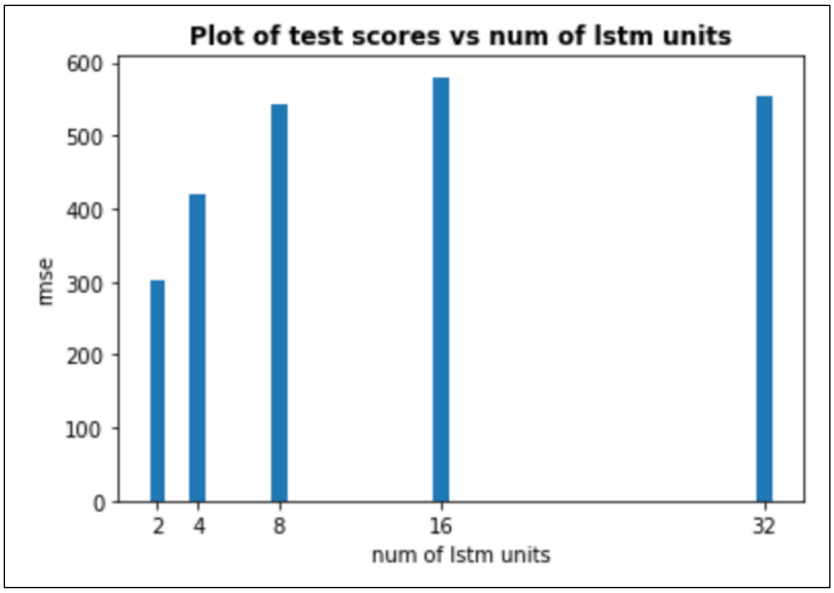
From correlation table of the 3 historical datasets below, weak correlation (21%) was observed between total grid demand and population, while strong correlation (88%) was between energy generated from solar panels and population. As such, population would be used as a feature for modelling of both total grid demand and energy generated from solar panels.

	grid_energy	population	solar_energy
grid_energy	1.000000	0.212893	0.034076
population	0.212893	1.000000	0.875866
solar_energy	0.034076	0.875866	1.000000

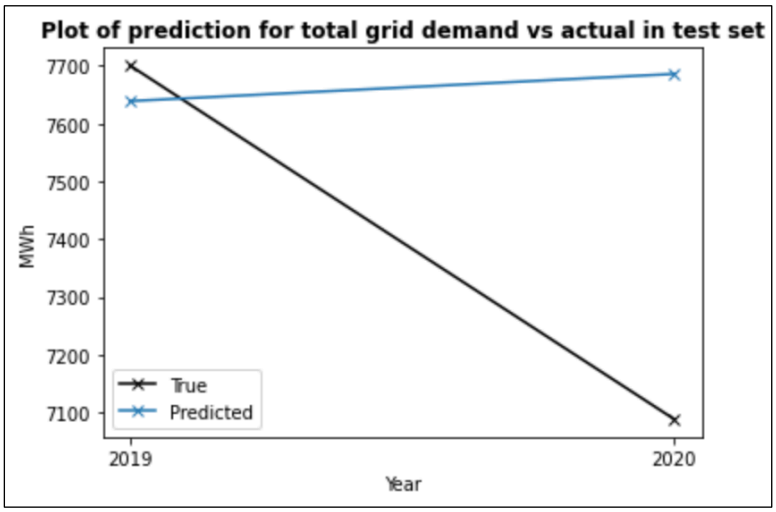
Modelling for total grid demand

Firstly, both population and total grid demand datasets were changed to yearly frequency and merged. Next, 4 features were created for modelling of total grid demand: year, population, lag 1 and lag 2 features. The merged dataset was then split into train and test set, in 80%/20% ratio. Data from train set was from 2012 to 2018, while data from test set was from 2019 to 2020. As train set was very small, we can expect the final model to not perform as well as when train set was sufficiently large. Next, normalizing of the dataset (to between 0 and 1) was carried out.

Long Short-Term Memory (LSTM) model was used for modelling. The graph below shows the root mean square error (RMSE) at different number of LSTM units. LSTM units of 2 had the lowest RMSE score, and subsequent modelling would be carried out with it.



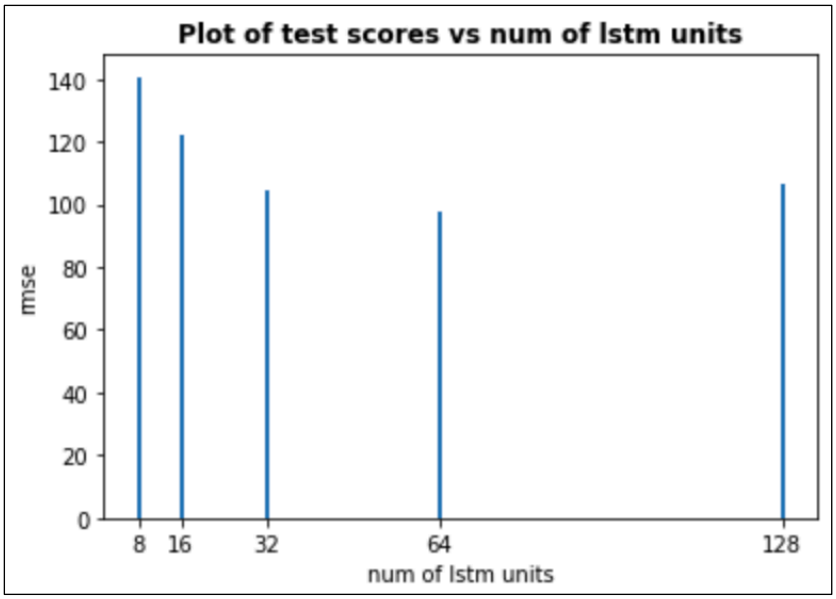
Below graph compares the predicted and true values in 2019 and 2020. Finally, the model was trained on the entire dataset (i.e. from 2012 to 2020) and will be used for forecasting into the future.



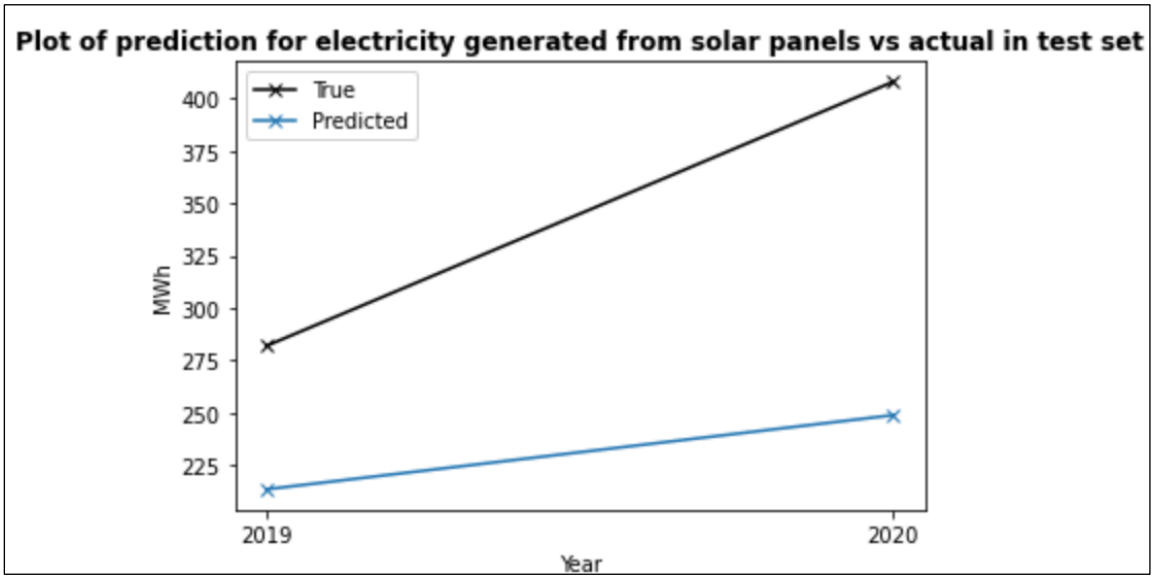
Modelling for energy generated from solar panels

Firstly, population dataset was changed to yearly frequency and merged with energy generated from solar panels dataset. Next, 4 features were created for modelling of total grid demand: year, population, lag 1 and lag 2 features. The merged dataset was then split into train and test set, in 80%/20% ratio. Data from train set was from 2011 to 2018, while data from test set was from 2019 to 2020. As train set was very small, we can expect the final model to not perform as well as when train set was sufficiently large. Next, normalizing of the dataset (to between 0 and 1) was carried out.

Long Short-Term Memory (LSTM) model was used for modelling. The graph below shows the root mean square error (RMSE) at different number of LSTM units. LSTM units of 64 had the lowest RMSE score, and subsequent modelling would be carried out with it.

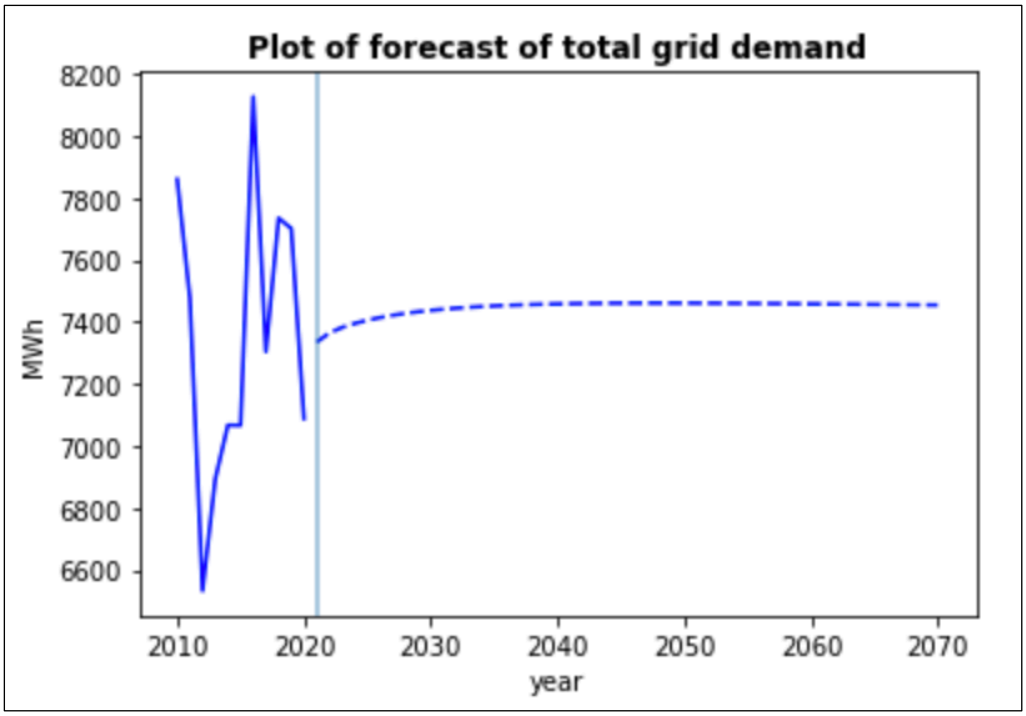


Below graph compares the predicted and true values in 2019 and 2020. Finally, the model was trained on the entire dataset (i.e. from 2011 to 2020) and will be used for forecasting into the future.

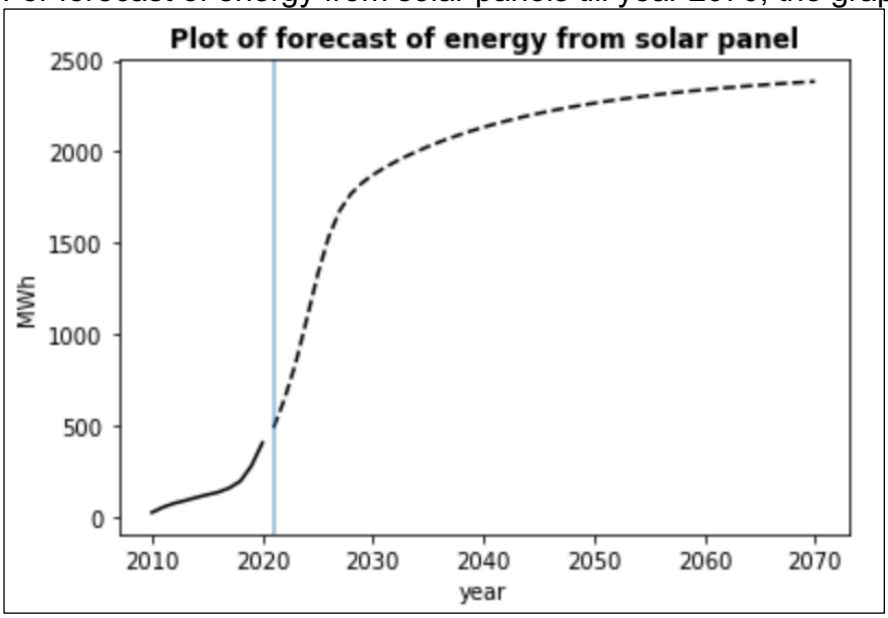


Forecasting

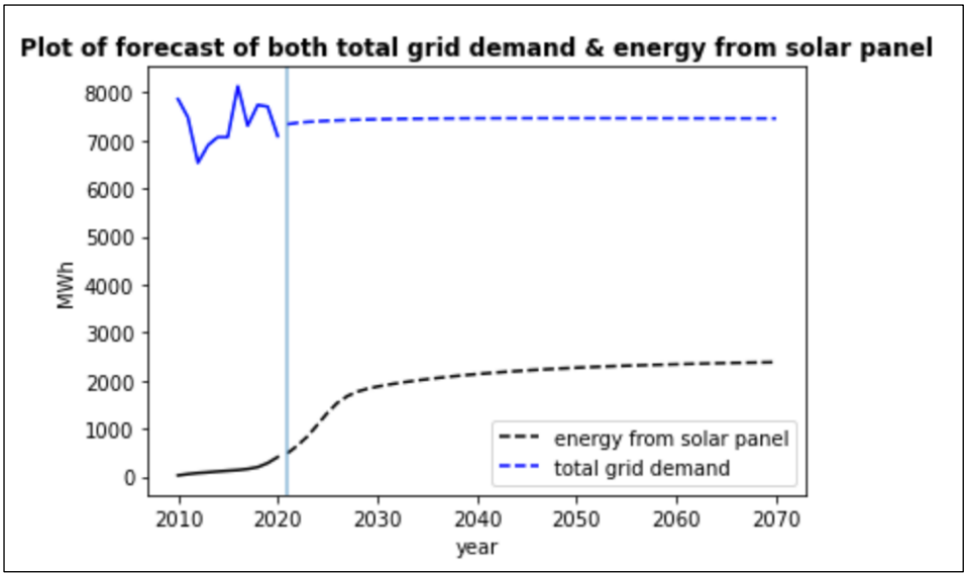
Forecasting for total grid demand and energy generated from solar panels were performed from 2021 to 2071 using projected population dataset as population feature. The graph below shows the forecast of total grid demand till year 2070. Due to small train dataset, only lag of 2 features were created to prevent further loss of data for training. Also, although population was used as a feature, it had a weak correlation to the target. The above factors could possibly explain why the model was unable to capture the fluctuations, as seen in historical data from 2009 to 2020.



For forecast of energy from solar panels till year 2070, the graph is as shown below.



Based on the modelling forecasts for both targets, unfortunately no intersection point was observed, as shown in graph below.



Conclusion

Forecasting of both total grid demand and energy generated from solar panels in NSW were performed in this project. Due to the small datasets available for training in both cases, it can be expected that the performance of both models to not perform well, especially for forecasting deep into the future. It is recommended to increase the size of time series datasets, as well as having more relevant features to improve the model.

Appendix 2 – Code

All our code can be found in the following Jupyter Notebooks in the Github repository:

EDA and Statistical Modelling: Group_Project_Report_Group_J.ipynb

ACF and PACF, and ML Modelling: ARIMA_REPORT_Rev01_U.ipynb