



CAPSTONE PROJECT BY TEAM D

A DATA SCIENCE APPROACH TO FORECASTING  
ELECTRICITY DEMAND IN NSW, AUSTRALIA

Baheerathan Gnanasundram, Matthew Seery, Mohammad Ahsan Ullah, Rahul Lobo.

School of Mathematics and Statistics  
UNSW Sydney

June 2021

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS OF  
THE CAPSTONE COURSE ZZSC9020

---

## Plagiarism statement

---

I declare that this thesis is my own work, except where acknowledged, and has not been submitted for academic credit elsewhere.

I acknowledge that the assessor of this thesis may, for the purpose of assessing it:

- Reproduce it and provide a copy to another member of the University; and/or,
- Communicate a copy of it to a plagiarism checking service (which may then retain a copy of it on its database for the purpose of future plagiarism checking).

I certify that I have read and understood the University Rules in respect of Student Academic Misconduct, and am aware of any potential plagiarism penalties which may apply.

By signing this declaration I am agreeing to the statements and conditions above.

Signed: \_\_\_\_\_ Date: \_\_\_\_\_

Signed: \_\_\_\_\_ Date: \_\_\_\_\_

Signed: \_\_\_\_\_ Date: \_\_\_\_\_

Signed: \_\_\_\_\_ Date: \_\_\_\_\_

---

## Acknowledgements

---

All thanks must go to our families and university professors who have guided us through this Capstone Project. Without them this would not be possible.

15/06/2021.

---

## Abstract

---

Forecasting electricity demand is an important requirement as there are now more interested stakeholders associated with the generation and distribution of this service. Traditionally electricity demand was just the concern of governments. However, this has now been extended to market bodies and owners and operators of the underlying infrastructure required for this service. Forecasting accurate energy demand not only supports network infrastructure, but also aides investment decisions about power generation. It is of thus of interest to a variety of stakeholders including power utilities, energy policymakers, and private investors just to name a few. This study aims to demonstrate how neural network (specifically LSTM neural networks) can be used to accurately forecast short-term energy demand. The focus will be on attributes such as temperature, month of the year, day of the week and time of the day as well as how past time lags can be used to predict future demand values. It is envisaged that the results of this study will provide useful insights for governments and market bodies to assist in the rules, policies and pricing that are invoked on the energy sector. Additionally, the businesses operating in this sector can use this information to guide their decision-making regarding electricity generation and distribution, as well as for the purpose of price benchmarking.

---

# Contents

---

Chapter 1	Introduction	1
Chapter 2	Literature Review	2
Chapter 3	Material and Methods	4
3.1	Software . . . . .	4
3.2	Description of the Data . . . . .	4
3.3	Pre-processing Steps . . . . .	5
3.4	Data Cleaning . . . . .	5
3.5	Assumptions . . . . .	6
3.6	Modelling Methods . . . . .	6
Chapter 4	Exploratory Data Analysis	7
4.1	Using R . . . . .	7
4.2	Using Python . . . . .	7
Chapter 5	Analysis and Results	9
5.1	A First Model . . . . .	9
Chapter 6	Discussion	10
Chapter 7	Conclusion and Further Issues	11
Appendix		14
<b>Codes</b>	. . . . .	14
<b>Tables</b>	. . . . .	14

---

# CHAPTER 1

## Introduction

---

Towards the end of the 20th century, countries throughout the world began moving away from regulated government-controlled energy supply to a deregulated sector influenced by market forces [1]. This means that forecasting for electricity demands is essential information required beyond previously regulated governmental structures. It is now required by the market bodies working in the sector and private industries who invest in the generation and distribution of electricity. In turn, this aids in better rules, policies, and pricing in the energy sector. However, for this study, the focus will primarily be on short-term forecasting as this can greatly aid vested parties concerned in maximising profits and cutting costs.

Temperature plays a significant role in electricity demand. This is because heating is utilised more by customers in the cooler months while cooling is required for the hottest months of the year. For that reason, temperature is an essential attribute that will be investigated in this study. However, temperature does not account for other factors such as humidity or varying electricity usage patterns associated with specific days of the weeks. For example, a temperature of 25 degrees in Spring may not lead to as much usage of air conditioning as a humid day in Summer with the same temperature. Additionally, the rate of usage of electricity will inevitably vary between weekdays, weekends, and public holidays due to the varying ratios of usage between business and residential customers. Therefore, the month of the year, day of the week and time of day will also be examined in this study.

---

## CHAPTER 2

### Literature Review

---

There have been various studies undertaken where more traditional statistical methods such as multiple linear regression (MLR) have been utilised to predict demand [2]. However, in more recent times there has been greater focus on the use of neural networks to help solve the problem of forecasting demand [3]. The advantage of neural networks is they are very suitable for determining non-linear relationships [3] and have been widely used for short-term demand forecasting [4]. They are also very flexible and easy to configure when dealing with time-series data [5]. For neural networks, monthly values have been a popular unit of measure and this is particularly true when short-term forecasting of demand has been employed [5]. However, this proposed solution differs because the focus will be on intervals shorter than even a day as this is of greater interest to the client. Nevertheless, the month of the year will still be considered as an input factor for the model for its potential to distinguish factors such as humidity that are not fully explained by temperature alone. Another distinguishing factor for this study will be the emphasis on the day of the week where each record will be categorised as either a weekday, weekend, or public holiday.

Up until the time of writing, the majority of deep learning models being applied to energy forecasting fall under the subset of three major ways [6]. These are: A feed forward neural network (FFNN)/Multi Layer Perceptron (MLP) through the process of increasing the number of hidden layers, some form of recurrence through a recurrent neural network (RNN), long-short term memory (LSTM) or gated recurrent unit (GRU), or through sequentially combining different types of algorithms into an overall structure. In 2020, Xue et al. contrasted these different approaches in order to forecast the heating demand of a district system [7]. Their experiments showed that the LSTM models were among the highest-performing models tested for.

When applied to natural gas forecasting, a close relative to electricity demand, RNN models have proven to be useful in prediction. Wei et al. (2019) explored the application of LSTM for forecasting natural gas consumption of four cities [7]. In addition, LSTM models used for forecasting were compared with other model techniques, including multiple linear regression, feed-forward neural networks, and a support vector regression in this study. The authors' rigorous testing demonstrated that LSTM models achieved a greater accuracy than the other data-driven models. It appears that RNN and LSTM models have formed the majority of accurate forecasting implementations to date, with other machine learning techniques such as SVMs performing as well as, but not better than RNN and LSTM based models.

An example of this is demonstrated by Amarasinghe et al. (2017), who contrasted the deep learning techniques of a Convolutional Neural Network (CNN) and LSTM, against a traditional machine learning technique (SVM) in order to forecast the energy demand of a building. This forecast was for a time period of sixty hours ahead and used 4 years of training data [8]. Their experiments proved that the deep learning based forecasting models obtained less forecasting error when compared with standard machine learning-based techniques; in particular, the LSTM obtained the smallest error.

Due to the significance of the LSTM accuracy across a variety of these studies, it will form the basis of this analysis.



---

## CHAPTER 3

### Material and Methods

---

#### 3.1 Software

A variety of software was used in the analysis as well as for collaboration and organisational purposes. GitHub was used as the main tool for collaboration and version control. It is not only extensively used in the industry but also enables the ability to collaborate on code seamlessly from local machines. This was the best choice as it is also used in the course instruction.

For data analysis purposes, Python, R/Rstudio and Tableau were all used. R was utilised in the initial exploratory data analysis phase in conjunction with Tableau, to visualise the dataset and to observe any obvious trends in the energy data. Tableau was especially useful in visualisation as it allowed the segmentation of days, times and months to be easily discerned in reference to energy demand. The main software used for the electricity demand modelling was Python. It was also used to clean, transform, and replace missing values in the supplied data and for merging separate datasets. including the final LSTM model which was constructed.

Furthermore, Excel was used to fill in missing values for the temperature dataset and the report for this analysis has been constructed in RMarkdown.

#### 3.2 Description of the Data

Three csv files were supplied for this project. The file `totaldemand_nsw.csv` contains three columns which record a date and timestamp (DATETIME) for the electricity consumed (TOTALDEMAND) for the associated region (REGIONID). The DATETIME columns contains dates between the 1st of January 2010 and midnight on the 18th of March 2021. Also included with each date is a timestamp with values for hours and minutes. There is a timestamp for each hour and thirty-minute interval within each hour for the aforementioned range of dates. All observations have the same value for the REGIONID column which is 'NSW1'.

The file `temperature_nsw.csv` also contains three columns. The columns record the location for the temperature observation (LOCATION), the date and time it was recorded (DATETIME) and the actual recorded temperature (TEMPERATURE). The values in the LOCATION column are all the same containing the value 'Bankstown' that describes the Bankstown weather station. The values in the DATETIME column are in the same format and range of dates as the `temperature_nsw.csv` file except some observations are outside the on hour or thirty-minute interval within each hour time periods. The file also contains duplicates and has

missing values between the dates of the 1st of January 2010 and 18th of March 2021.

The third file, `forecastdemand_nsw.csv`, contains periodic forecasts (FORECAST-DEMAND) for a set date and time (DATETIME) and the date and time the forecast was made (LASTCHANGED). The values in the FORECASTDEMAND column align with the same range of dates in the `totaldemand_nsw.csv` file. The format of the date and time in this column and the DATETIME column is the same as the DATETIME column in the `totaldemand_nsw.csv` file. The `forecastdemand_nsw.csv` file also includes a REGIONID column and all with the same values as the `totaldemand_nsw.csv` file. There is also a column called PREDISPATCHSEQNO which contains a unique id number for each forecast and a PERIODID column which contains a unique ID for forecasts of each unique date and time in the DATETIME column. The data in this file was not used in the neural network models that were tested.

### 3.3 Pre-processing Steps

What did you have to do to transform the data so that they become useable?

### 3.4 Data Cleaning

The data in the file `totaldemand_nsw.csv` required no further cleaning other than the removal of the REGIONID column as it contains redundant data. For the `temperature_nsw.csv` file, further cleaning was required. There were 14 duplicates found which were removed from the dataset. There were also 579 records that were either missing or had date and timestamps that did not align with those in the `totaldemand_nsw.csv` file. An attempt was made to reassign the inconsistent timestamps to the nearest 30-minute interval if one for that period did not already exist. However, this only reduced the number of incomplete records down to 564. Instead, it was decided to fill in the missing values by sourcing these temperatures elsewhere.

For dates between March 2016 and April 2018, hourly temperature recordings were collected for the Bankstown weather station (Station Number 66137) from datasets found at [data.gov.au](http://data.gov.au) [9]. Between May 2010 and May 2011, there were sections of the dataset where more than 10 consecutive timestamps were missing. Additionally, they passed through time periods where the maximum or minimum temperature of the day normally occurred. Therefore, either a maximum or minimum temperature for these days was sourced from the Bureau of Meteorology [10] for the Bankstown weather station. These temperatures were then assigned to an estimated hour of the day where the maximum or minimum temperature was likely to have occurred based on when maximum or minimum temperatures occurred on the 3 preceding and following 3 days. There was also missing timestamps for the 21st of May 2018 between 10:30 and 17:00. However, not even a maximum temperature could be sourced for this day. Therefore, an estimate for the maximum temperature and the time of day it occurred was made using the preceding and following 3 days as a guide. These sourced temperatures and their associated

timestamps were placed in a separate csv file and then merged with the original temperature dataset.

All remaining missing temperatures were filled using the resample and interpolate functions associated with panda dataframes in python. This assigned incremental and evenly spaced values ranging between the values of the previous and next existing temperatures that surrounded the consecutive missing values. Prior to the merging of the two datasets, the LOCATION column was dropped from the total demand dataset as all the values were the same and thus redundant.

### 3.5 Assumptions

What assumptions are you making on the data?

### 3.6 Modelling Methods

---

## CHAPTER 4

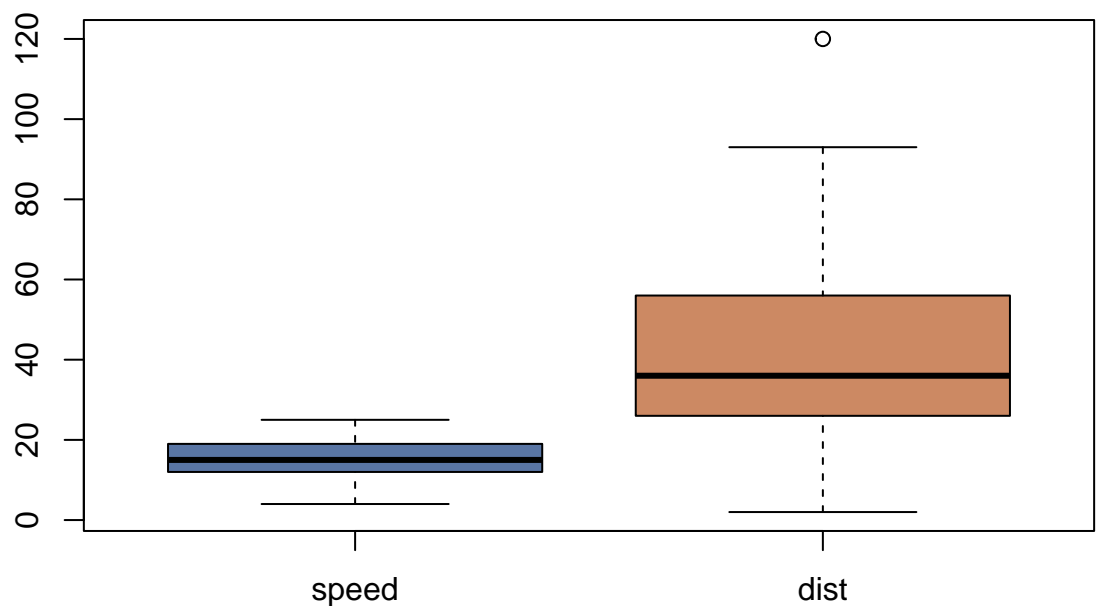
### Exploratory Data Analysis

---

This is where you explore your data using histograms, scatterplots, boxplots, numerical summaries, etc.

#### 4.1 Using R

```
boxplot(cars, col = c("#5975a4", "#cc8963"))
```



#### 4.2 Using Python

See [https://cran.r-project.org/web/packages/reticulate/vignettes/r\\_markdown.html](https://cran.r-project.org/web/packages/reticulate/vignettes/r_markdown.html) for more details.

You need to install the R package `reticulate`.

```
print("Python can be used with MATHxxxx!")
```

```
## Python can be used with MATHxxxx!
```

```
import sys
print(sys.version)
```

```
## 3.6.12 |Anaconda, Inc.| (default, Sep  9 2020, 00:29:25) [MSC v.1916 64 bit
```

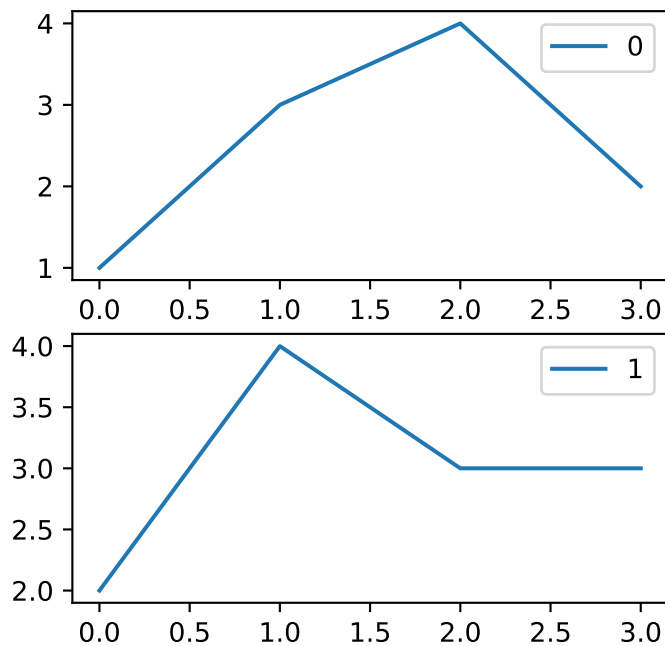
```
import numpy as np
np.random.seed(1)
np.random.normal(0.0, 1.0, size=10)
```

```
## array([ 1.62434536, -0.61175641, -0.52817175, -1.07296862,  0.86540763,
##        -2.3015387 ,  1.74481176, -0.7612069 ,  0.3190391 , -0.24937038])
```

```
import pandas as pd
import matplotlib.pyplot as plt
df=pd.DataFrame([[1, 2], [3, 4], [4, 3], [2, 3]])
fig = plt.figure(figsize=(4, 4))
for i in df.columns:
    ax=plt.subplot(2,1,i+1)
    df[[i]].plot(ax=ax)
    print(i)
```

```
## <AxesSubplot:>
## 0
## <AxesSubplot:>
## 1
```

```
plt.show()
```



---

## CHAPTER 5

### Analysis and Results

---

#### 5.1 A First Model

Having a very simple model is always good so that you can benchmark any result you would obtain with a more elaborate model.

For example, one can use the linear regression model

$$Y_i = \beta_0 + \beta_1 x_{1i} + \cdots \beta_p x_{pi} + \epsilon_i, \quad i = 1, \dots, n.$$

where it is assumed that the  $\epsilon_i$ 's are i.i.d.  $N(0, 1)$ .

---

## CHAPTER 6

### Discussion

---

Put the results you got in the previous chapter in perspective with respect to the problem studied.

---

## CHAPTER 7

### Conclusion and Further Issues

---

What are the main conclusions? What are your recommendations for the “client”?  
What further analysis could be done in the future?

A figure:



Figure 7.1: A caption

In the text, see Figure [7.1](#).

---

# References

---



---

## References

---

- [1] J. Catalão, S. Mariano, V. Mendes, L. Ferreira, [Short-term electricity prices forecasting in a competitive market: A neural network approach](#), *Electric Power Systems Research* 77 (10) (2007) 1297–1304. doi:<https://doi.org/10.1016/j.epsr.2006.09.022>.  
URL <https://www.sciencedirect.com/science/article/pii/S0378779606002422>
- [2] Z. Mohamed, P. Bodger, [Forecasting electricity consumption in new zealand using economic and demographic variables](#), *Energy* 30 (10) (2005) 1833–1843. doi:<https://doi.org/10.1016/j.energy.2004.08.012>.  
URL <https://www.sciencedirect.com/science/article/pii/S0360544204003639>
- [3] E. González-Romera, M. Jaramillo-Morán, D. Carmona-Fernández, [Monthly electric energy demand forecasting with neural networks and fourier series](#), *Energy Conversion and Management* 49 (11) (2008) 3135–3142, special Issue 3rd International Conference on Thermal Engineering: Theory and Applications. doi:<https://doi.org/10.1016/j.enconman.2008.06.004>.  
URL <https://www.sciencedirect.com/science/article/pii/S0196890408002288>
- [4] G. Ciulla, A. D’Amico, [Building energy performance forecasting: A multiple linear regression approach](#), *Applied Energy* 253 (2019) 113500. doi:<https://doi.org/10.1016/j.apenergy.2019.113500>.  
URL <https://www.sciencedirect.com/science/article/pii/S0306261919311742>
- [5] E. G. D. Carmona, M.A. Jaramillo, J. Alvarez, [Electric energy demand forecasting with neural networks](#).  
URL <https://www.sciencedirect.com/science/article/abs/pii/S0301421595001166>
- [6] R. Kumar, R. Aggarwal, J. Sharma, [Energy analysis of a building using artificial neural network: A review](#), *Energy and Buildings* 65 (2013) 352–358. doi:[10.1016/j.enbuild.2013.06.007](https://doi.org/10.1016/j.enbuild.2013.06.007).
- [7] G. Xue, C. Qi, H. Li, X. Kong, J. Song, [Heating load prediction based on attention long short term memory: A case study of xingtai](#), *Energy* 203 (2020) 117846. doi:[10.1016/j.energy.2020.117846](https://doi.org/10.1016/j.energy.2020.117846).
- [8] D. L. Marino, K. Amarasinghe, M. Manic, [Building energy load forecasting using deep neural networks](#) (Oct 2016).  
URL <https://arxiv.org/abs/1610.09460>
- [9] [\[link\]](#).  
URL <https://data.gov.au/search>

- [10] [\[link\]](#).  
URL <http://www.bom.gov.au/climate/data/stations/>

---

## Appendix

---

### Codes

Add you codes here.

### Tables

If you have tables, you can add them here.

Use [https://www.tablesgenerator.com/markdown\\_tables](https://www.tablesgenerator.com/markdown_tables) to crete very simple markdown tables, otherwise use L<sup>A</sup>T<sub>E</sub>X.

Tables	Are	Cool
col 1 is	left-aligned	\$1600
col 2 is	centered	\$12
col 3 is	right-aligned	\$1