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 and Data Cleaning . . . 5
- 3.4 Assumptions . . . 6
- 3.5 Modelling Methods . . . 6
- 3.6 Prediction Algorithm . . . 8

Chapter 4 Exploratory Data Analysis 9
- 4.1 Using Tableau . . . 9

Chapter 5 Analysis and Results 17
- 5.1 A First Model . . . 17

Chapter 6 Discussion 18

Chapter 7 Conclusion and Further Issues 19

Appendix 22
- **Codes** . . . 22
- **Tables** . . . 22

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 appliied 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 Amarasignhe 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 (FORECASTDEMAND) 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 and 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 [9]. Between May 2010 and May 2011, there were sections of the dataset where more than10 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 proceeding 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 proceeding 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 temperature dataset as all the values were the same and thus redundant.

Using python, the temperature dataset was also extended with new features using the DATETIME column as a base. First, separate columns were made for day, day of the week, month, year, hour and minute. Most columns maintained the same values as represented in the original date and timestamp. The only exception was the day of the week which used a numbering scale from 0 (Monday) to 6 (Sunday) to represent the days of the week. The values in the day of week column were then used to generate a Boolean column where 0 signifies a weekday and 1 represents a weekend.

Additional columns were also generated using the values in the HOUR and MONTH columns. The 0 to 23 hours in the HOUR column were broken up into 4 sectors where 0 (Midnight) to 5 represents early morning, 6 to 11 represents late morning, 12 (Midday) to 17 represents afternoon and 18 to 23 represents evening. These grouping were given values from 1 to 4 respectively. Using the MONTH column, a column called SEASON was also created with 1 representing summer, 2 representing autumn, 3 representing winter and 4 representing spring. The temperature dataset was then merged with the total demand dataset. After the transformed dataset was saved as a csv file, one additional column was added in Microsoft Excel called PUBLIC_HOLIDAY with a value of 1 given if the date was a public holiday and 0 otherwise. See Fig. 3.1 for the final processed dataset used in the modelling.

DATETIME	TEMPERATURE	DAY	DAY_OF_WEEK	MONTH	YEAR	HOUR	MINUTE	WEEKEND	PUBLIC_HOLIDAY	TIME_OF_DAY	SEASON	TOTALDEMAND
1/01/2010 0:00	23.1	1	4	1	2010	0	0	0	1	1	1	8038
1/01/2010 0:30	22.9	1	4	1	2010	0	30	0	1	1	1	7809.31
1/01/2010 1:00	22.6	1	4	1	2010	1	0	0	1	1	1	7483.69
1/01/2010 1:30	22.5	1	4	1	2010	1	30	0	1	1	1	7117.23
1/01/2010 2:00	22.5	1	4	1	2010	2	0	0	1	1	1	6812.03
1/01/2010 2:30	22.4	1	4	1	2010	2	30	0	1	1	1	6544.33
1/01/2010 3:00	22.3	1	4	1	2010	3	0	0	1	1	1	6377.32
1/01/2010 3:30	22.3	1	4	1	2010	3	30	0	1	1	1	6282.85
1/01/2010 4:00	22.1	1	4	1	2010	4	0	0	1	1	1	6211.49
1/01/2010 4:30	22.2	1	4	1	2010	4	30	0	1	1	1	6248.31

Figure 3.1: The first 10 records in the transformed dataset used for modelling of the neural networks.

3.4 Assumptions

What assumptions are you making on the data?

3.5 Modelling Methods

Before the dataset described in Chapter 3.3 could be used for modelling, transformations were required for some attributes. Columns with ordinal number values, including the columns DAY, DAY_OF_WEEK, MONTH, HOUR, TIME_OF_DAY

and SEASON, were transformed using one-hot encoding. The dataset was then split into a training and test set. A minimum maximum scalar was then fitted for each of continuous values in the training set. This was performed on the TEMPERATURE and TOTALDEMAND columns where values were proportionally reassigned to values between 0 and 1. These fittings were then used to transform the TEMPERATURE and TOTALDEMAND columns in both the training and test set. By Experiment we limit the inputs to month and hour as categorical variables, identification as whether it is a weekday or weekend and the day happened to be a public holiday.

The LSTM model has been designed to use the attributes of a designated number of immediately previous timestamps to the record that is either being trained or predicted. This number is set by a variable called 'time_steps'. A function was then created to transform both the X value for the training and test datasets into a three-dimensional matrix. The three dimensions represent the number of records in the dataset, the number of immediately prior records (time_steps) to the current record and the number of features used for these prior records. This therefore means that for each dataset the first number of records, equal to the value of the time_steps variable, cannot be trained or predicted by the model. This is because the number of previous records in the dataset for these records is less than the value of the time_steps variable. It should also be noted that the value of the total demand for the previous records were also included in the matrix of X values. The y value fed to the neural network is the total demand for the current record being either trained or predicted.

The neural network consists of a single layer bidirectional LSTM with 50 neurons which uses 'relu' for the activation function. This feeds into a single dense neuron. As part of the python tensorflow library, keras was used to compile the model. Mean square error was used to measure the performance of the model and Adam was used to optimise the model. When fitting the model, 30 epochs were used with a batch size of 32 with 0.1 (10%) of the training dataset set aside for validation. The values of the test dataset were then used to make predictions for the y values of the test dataset with an inverse transform then applied to restore these predictions back to the pre-scaled values. Root Mean Square Error (RMSE) was also used to determine how well the model was performing.

The test result was recorded along with the 'DATETIME' timestamp, temperature, actual and predicted demand in a .csv file for further analysis and visualisation by any appropriate tools such as Tableau.

The built model and the scalers of temperature and demand were saved so that the prediction program can work independently by loading this model and scalers. Various combinations of the attributes were used until the best combination was determined. The final columns used in the model were TEMPERATURE, WEEKEND, PUBLIC_HOLIDAY, MONTH (one-hot encoded), HOUR(one-hot encoded) and TOTALDEMAND.

3.6 Prediction Algorithm

Based on the model, the forecast temperature is required to forecast demand. The data for 'time_steps' (e.g., 24) number of records of actual demand prior to the forecast period also required for the time series LSTM model to work. All other inputs can be prepared based on the 'DATETIME' timestamp.

Suppose the actual demand information available weekly on Sunday midday, and the requirement is to provide a forecast of demand for the following week. With the forecast of temperature for that week, the input data will be built with the appropriate timeslot related input, temperature, with the actual demand filed set to zero.

The last 24 timeslots of complete data records including actual demand will be joined in front of the forecast records. The prediction program loads the saved neural network model, scalers for temperature and demand and load the entire dataset and convert to appropriate input format with the necessary scaling. As the month and hour data (if the forecast only for few hours, not for 24 hours or a week) will be limited, it can not use standard one hot encoding. So, the program builds this encoding programmatically.

Once the input is transformed, the algorithm, picks the first 24 records, and predict the demand for the 25th slot. While recording this information for output, the actual demand for the 25th record will be updated with this predicted demand value and the oldest record (1st record) will be dropped.

Now having complete records for the first 24 records again, the above process is repeated until all the necessary forecast demands having been performed. Practically the future forecast demand is based on the forecast, so it is possible to have the error build cumulatively. However, if the actual demand data is available say every hour, then the program can predict the forecast hourly with much accurate input. The forecast period also will be decided by the stakeholders.

The test result was recorded along with the 'DATETIME' time stamp, temperature, predicted demand in a .csv file for distribution to relevant stakeholders. The model can be re-trained every month or even daily to cover latest environmental and policy impacts as training will take less than 10 minutes on an average computer without the GPU support for 3 year data.

Chapter 4

Exploratory Data Analysis

4.1 Using Tableau

Year to year monthly temperature-demand trend

- Given temperature and demand data is from Jan'10 to Mar'21
- Monthly avg. temperature varies 2/3 degrees
- Demand is comparatively higher during winter season (Jun-Aug)
- Overall demand shows reduction in the recent years

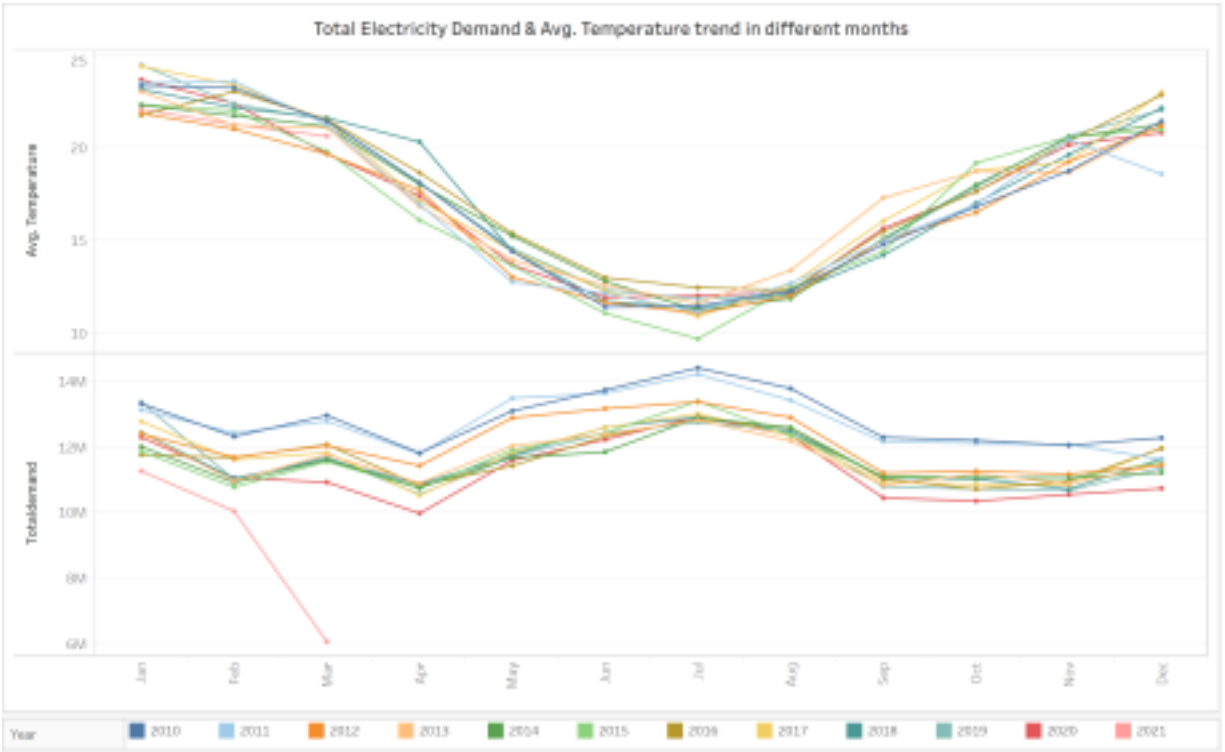

Figure 4.1: Total Electricity Demand and Avg. Temperature trend in different months

Monthly temperature-demand in diff. weekdays trend

- Different months have different demand pattern in different weeks days

- Year to year the pattern is not same

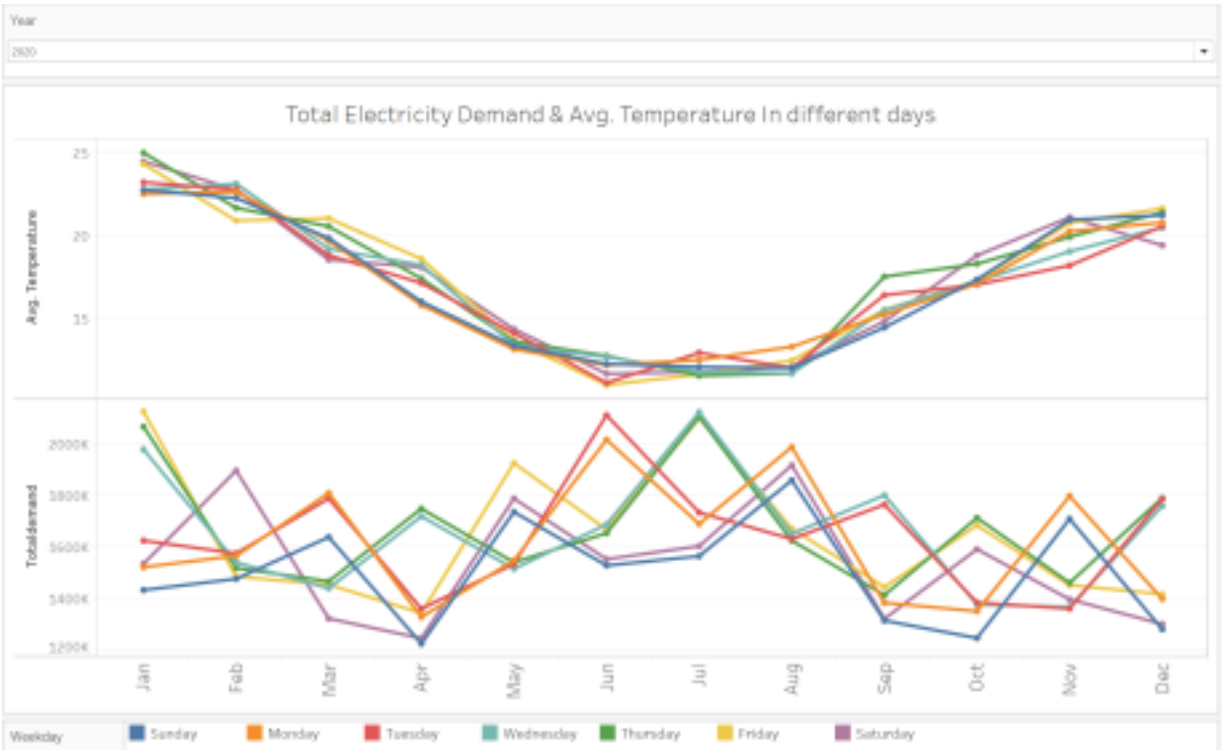

Figure 4.2: Total Electricity Demand and Avg. Temperature trend in different days

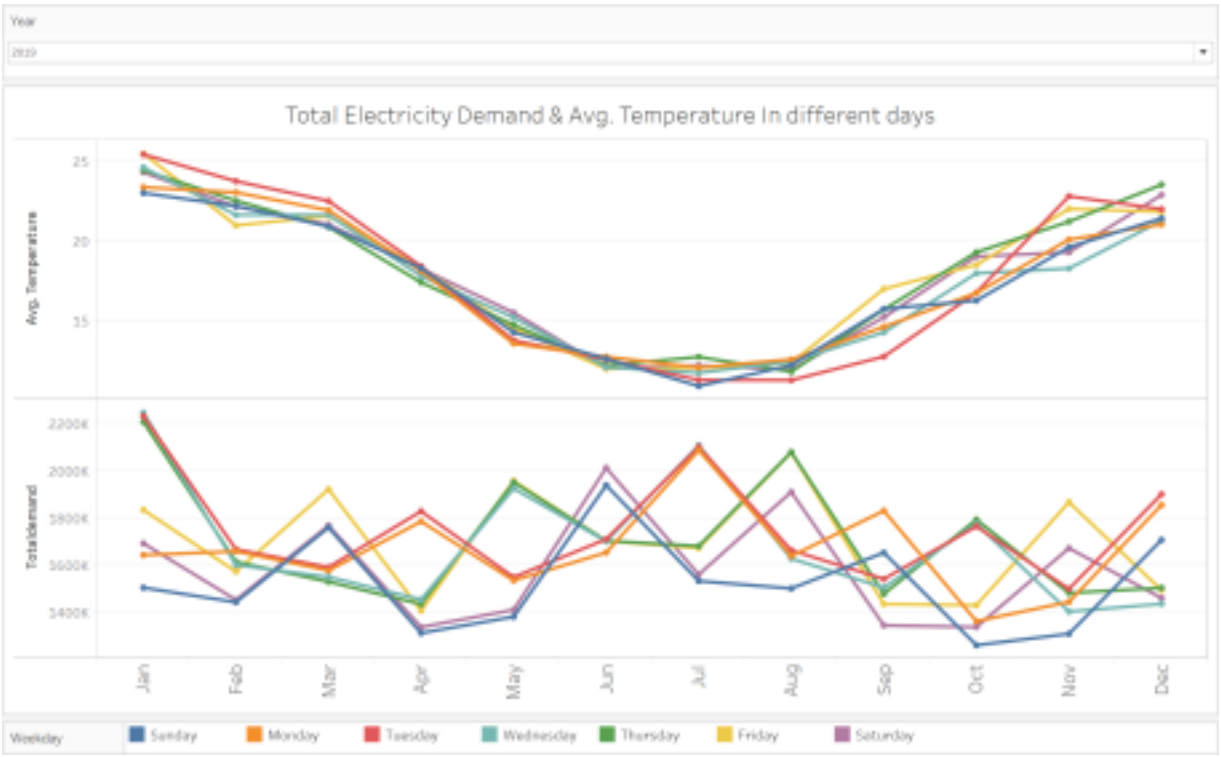

Figure 4.3: Total Electricity Demand and Avg. Temperature trend in different days

Hourly temperature-demand in diff. weekdays trend

- Demand somewhat follows temperature during off peak (night-time) time
- Demand variation during daytime varies in different days and not always proportional to temperature

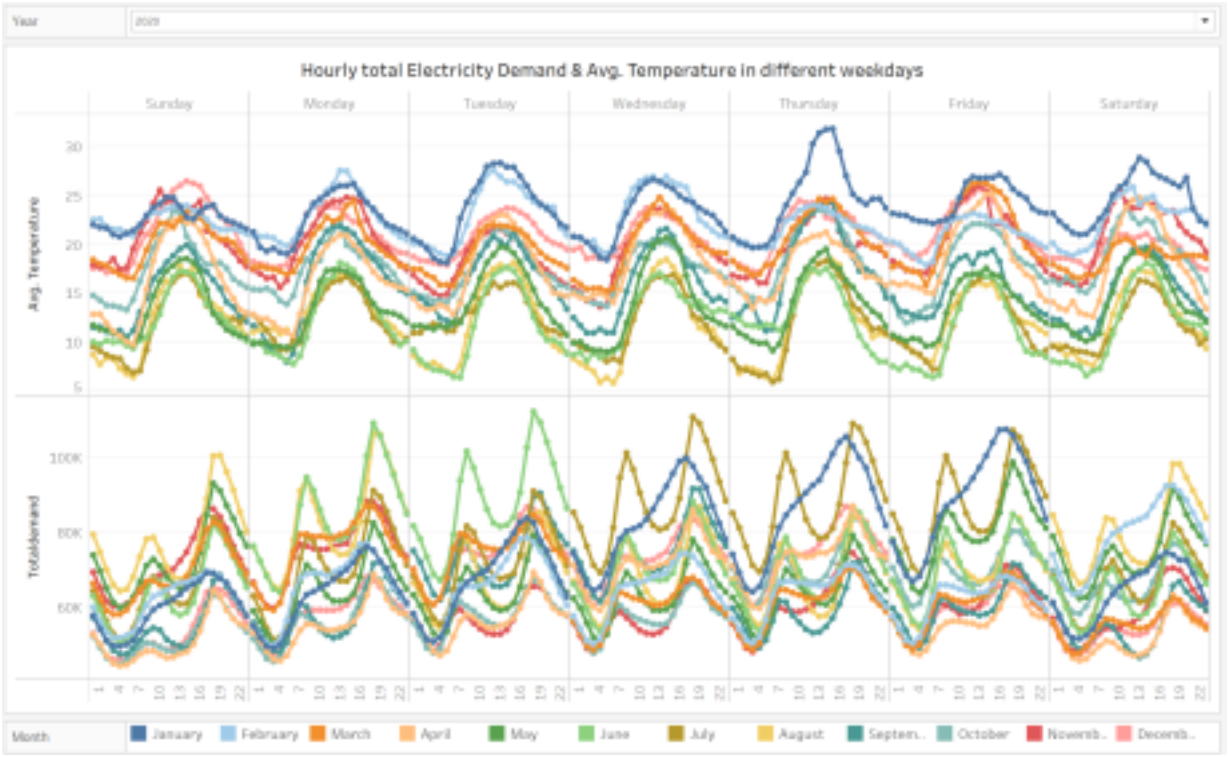

Figure 4.4: Hourly total Electricity Demand and Avg. Temperature in different weekdays

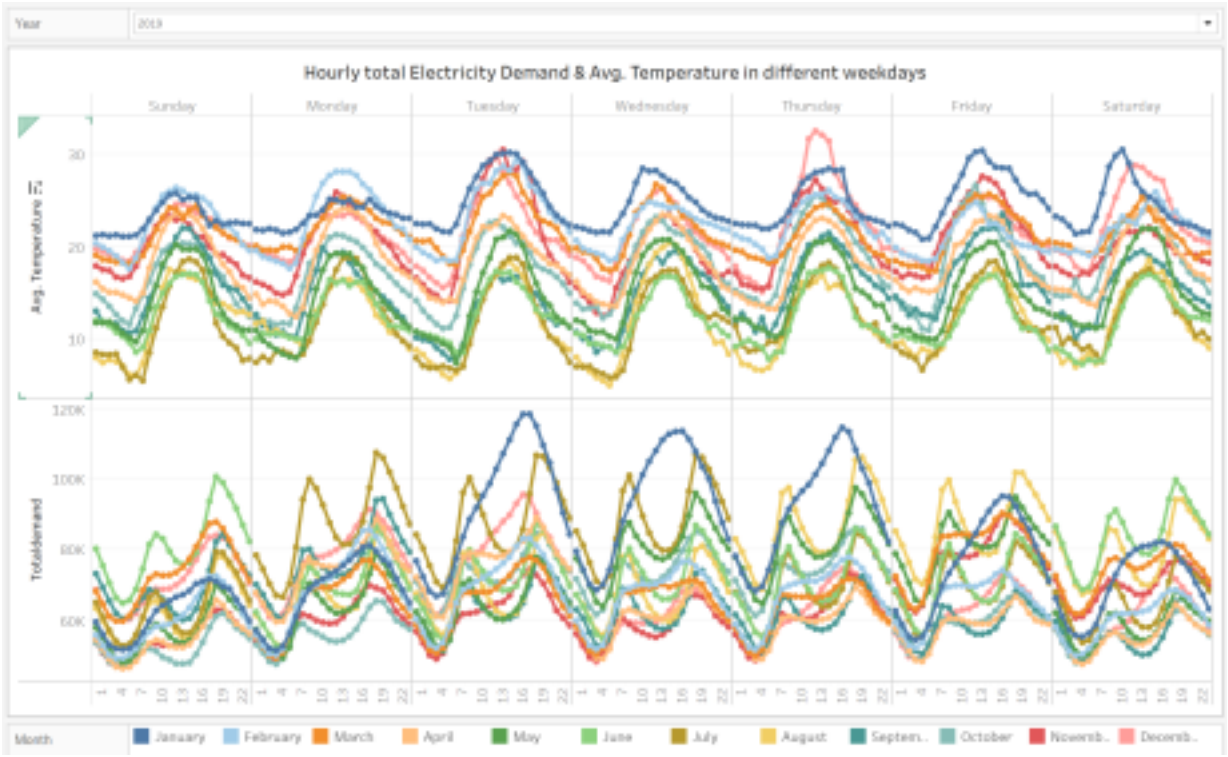

Figure 4.5: Hourly total Electricity Demand and Avg. Temperature in different weekdays

Hourly Forecast vs. Demand (Jan and Aug 2020 first 10 days)

- Current forecasts are pretty good in some hours and in some months
- Forecast suffers accuracy during daytime

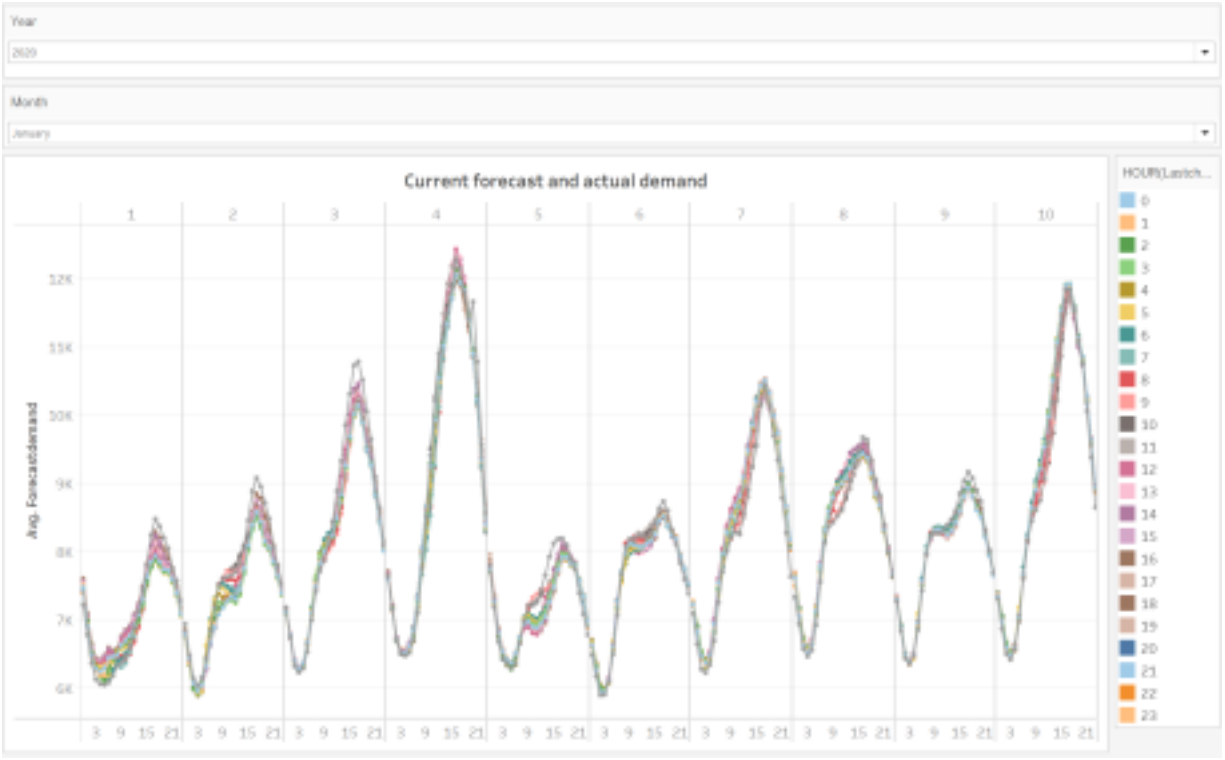

Figure 4.6: Current forecast and actual demand

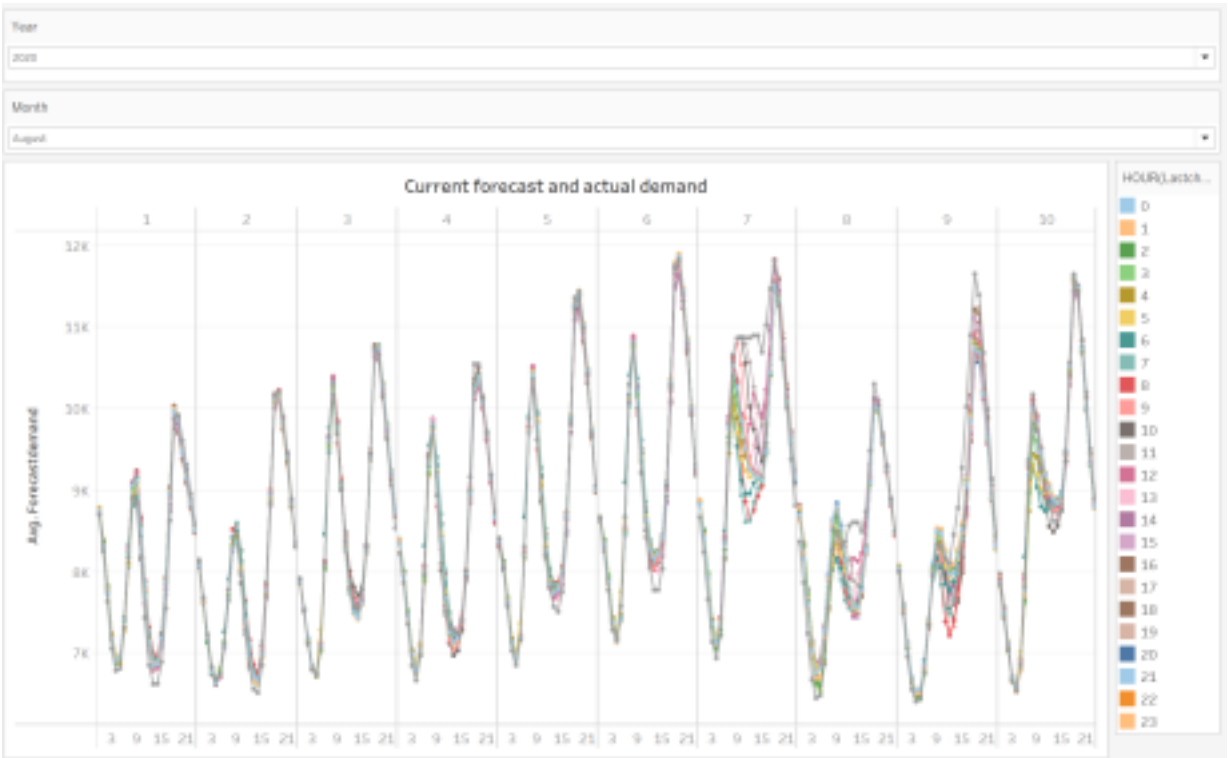

Figure 4.7: Current forecast and actual demand

2021 prediction results vs actual demand

- LSTM model output shows good performance during off peak time
- In some month (March) prediction seems very good

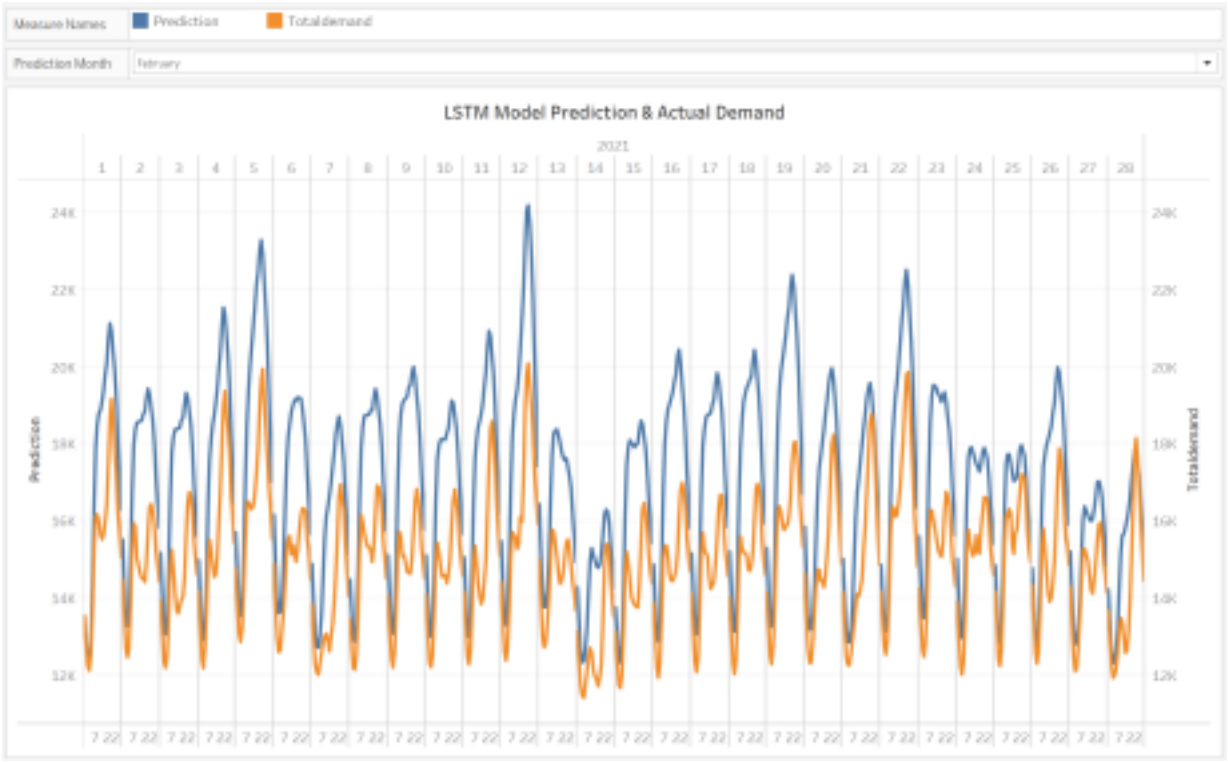

Figure 4.8: LSTM Model Prediction and Actual Demand

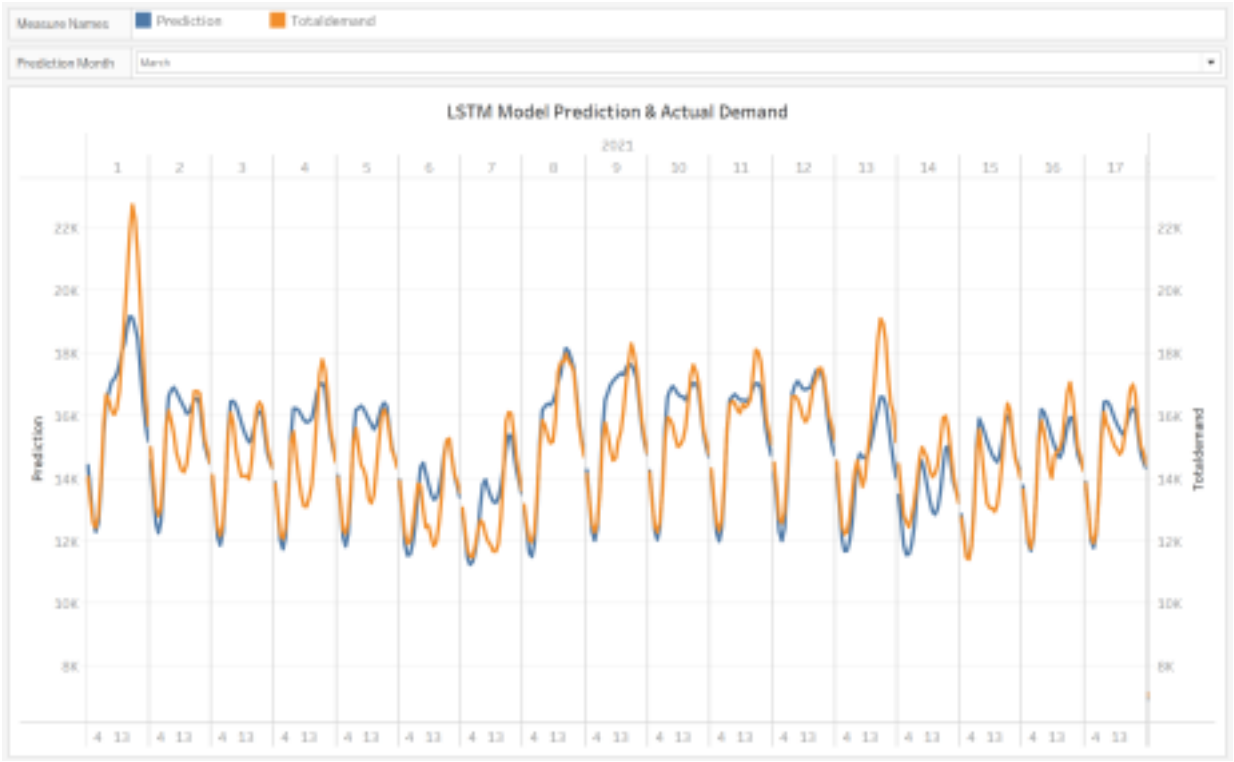

Figure 4.9: LSTM Model Prediction and Actual Demand

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

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

where it is assumed that the ϵ_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? The Long short-term memory (LSTM) model performs well when predicting electricity demand for NSW, Australia. Using the last X timestamps, the pre-processed data is given as input which is then processed by the LSTM and gives a Root Mean Square Error (RMSE) of X. This can be used profitably by the client through the appropriate pricing based on model outputs, with an appropriate profit margin to account for variance in the model. Although the LSTM has worked well, there is no doubt that more data (i.e. Solar PV data or even an increased granularity in timestamps) could decrease the model error and make pricing more efficient for the client. Furthermore, other deep-learning architectures may be explored as well, even though an LSTM worked successfully on this occasion. However, it is important to note that, as was evident in this analysis, that inducing a more complex network structure to model the data did not always improve model accuracy.

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/0301421595001166

[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.

[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.

[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 to crete very simple markdown tables, otherwise use LaTeX.

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