A Data Science Approach to Short-Term Electricity Demand Forecasting

Louis Long (z5162359) Mark Budrewicz (z5353932) Hao Chiou (z5131909)
Charlet Jeyapaul (z5375906)

07/10/2023

Contents

1 Abstract

2 Motivation

Electricity demand forecasting is a topic that is within the interest of government and market bodies, with the consensus of accurate forecasts of energy demand driving profitability. Electricity in Australia is generated primarily using oil, coal, and gas as fuel(Bhattacharya, Inekwe, and Sadorsky 2020). Also, with oil prices increasing, this implies that the cost to generate energy will also increase(Kpodar and Liu 2022). Therefore, profit maximization is achieved when energy production is minimized to be near actual energy demand. Thus, if electricity demand can be modelled accurately, government and market bodies can then utilize such information to tailor policies as well as optimize pricing. A recent phenomenon in the electricity grid has been the shift towards renewable energy sources such solar photovoltaic and wind generation. While these technologies are indeed cleaner they bring with them risks(Australian Renewable Energy Agency 2021) due to their intermittency and inability to be dispatchable as compared to electricity produced from fossil fuels. As of 2021, renewables accounted for 23% of total electricity generation in the state of NSW(Department of Climate Change, Energy, the Environment and Water 2021). As it is non dispatchable, renewable electricity generation acts to lower the energy demand of the grid, so much so, that the Australian Energy Market Operation (AEMO) has already acknowledged that the increasing uptake in rooftop solar will shift maximum load demand to later in the day at sunset as temperatures remain high but roof top solar output falls to nil(Australian Energy Market Operator 2017). The increase in renewable electricity generation will see a parallel requirement for flexible electricity generation such as gas, pumped hydro and grid scale batteries with quick ramp up ability to cope with the variability in supply(AGL Energy 2020). And with that, a more robust shorter term electricity demand forecasting process will be required to ensure that the quick start alternative electricity generators have sufficient warning to ramp up when required.

3 Introduction

The aim of this paper is to investigate different modelling techniques to generate a rolling 60 min forward demand forecast for the state NSW electricity grid that could be used by quick start pumped hydro electricity generator to help minimise operational costs and give a competitive edge in pricing which will benefit its profitability and flow down to lower costs for the end users of the electricity grid.

Quick start electricity generators such as pumped hydro are essential to the electricity generation mix as they add flexibility to the grid to cope with highly variable demand (AGL Energy 2020). This flexibility ensures a sufficient balance of supply and demand as these quick start generators can supply with a shorter ramp up period. While the ramp up period is shorter than conventional base load generators, it still important for an electricity generator to be able to estimate how much electricity is required by the grid within a short forecast window so that it can strategically plan generation output to minimise its startup and shut down costs and maximise profit. Times of high load requirement in the grid inevitable lead to higher spot prices and conversely times of low load lead to lower spot prices. Pumped hydro generators act like large scale batteries for the grid with a unique position to take advantage of both ends of the demand spectrum, dispatching electricity at times of high load, and consuming electricity from the grid at times of low demand to pump water back upstream to "recharge" their water reservoirs. So, it is incumbent on pumped hydro operators to be able to accurately forecast and capitalise on such situations, in line with its operational constraints (startup and shutdown times, maintenance requirements). While an electricity demand forecast is already generated by AEMO (Australian Energy Market Operator 2023b) for participants in the grid, this analysis aims to investigate a forecasting model that will outperform AEMO's demand forecast, allowing a supplier to take advantage of potential misalignment in AEMO's forecast to actual load demand.

AEMO produces a short term forecast model in 30 minute intervals (the 'Half-hourly demand model') (Australian Energy Market Operator 2023b), that uses features such as a categorical value for each half hour block of the day and month of the year, dummy variables for weekends and public holidays and transformations of half hourly temperature inclusive of interaction terms. This model uses a Least Absolute Shrinkage and Selection Operator (LASSO) regularisation algorithm which is a regression analysis method

that automatically performs feature selection by including a penalty term to the loss function which acts to shrink the non significant variable coefficients to zero.

Our models seek to build upon the AEMO 'Half-hourly demand model', to test if other meteorological data such as solar exposure can capture the impact of household solar photovoltaic generation in the grid, and whether rainfall has an impact on indoor electricity consumption. Our investigation seeks to compare the performance various time series and machine learning models such as LASSO, Autoregressive integrated moving average (ARIMA), Support Vector Regression (SVR) and Random Forrest Regression with the aim of improving the short term electricity demand forecast with a slight increase in computational cost. Each of the models will then be further refined via training with simulated data. The final model will be selected after comparison of various generated models and features, specifically through consideration of feature selection, accuracy, and interpretability and performance relative to AEMO's forecast.

4 Literature review

Initial review of the literature on electricity demand forecasting pointed to a well understood relationship between temperature and electricity demand that has already been incorporated by the AEMO's forecasting procedures (Australian Energy Market Operator 2023b). This relationship has been widely explored with a specific focus on higher frequency time intervals. McCulloch and Ignatieva (2020) showed a strong relationship when modelling Australian 5-minute electricity demand with intraday temperature using a Generalised Additive Model. This same research also indicated that sensitivity of demand to temperature was heavily dependent on the time of day with higher demand sensitivity during hours of high human activity. In G. Zhang and Guo (2020), their support vector machine (SVM) model found other meteorological features such as wind speed, relative humidity, precipitation and air pressure impacted electricity demand. Interestingly the behaviour of these variables on demand was non linear, which the authors hypothesised could be due to their impact on relative temperature (ie high wind can act to help cool on warmer days, but make it feel colder on cooler days). 'Vu, Muttaqi, and Agalgaonkar (2015)'s regression model for monthly electricity demand for the State of NSW showed that demand was "predominantly dependent on temperature, humidity and the number of rainy days". While the study was done on monthly electricity demand data, it points to a link in meteorological features impacting NSW electricity demand which may also transfer to more granular time periods.

Other literature included additional features and characteristics of the data such as seasonality and cyclicality (from hourly to monthly) in the electricity demand time series. As described in Australian Energy Market Operator (2023b), AEMO's half hourly model includes categorical values for half hourly block of the day and month of the year, and include dummy variables for non work days. In Clements, Hurn, and Li (2016) study on Queensland electricity demand, seasonality of demand including hour of the day, day of the week were identified as important features that needed to be factored into the modelling. One approach by Jiang, Li, Liu, et al. (2020) was to remove the seasonality as a prepossessing step through a Fast Fourier Transform method which converted the electricity demand from time series domain to frequency domain. In this model, the seasonality was removed so as not to affect the SVM algorithm. Other approaches such as Alonso, Nogales, and Ruiz (2020), chose to let their Long Short-Term Memory (LSTM) recurrent neural network to capture the seasonal idiosyncrasy naturally. Through their initial data exploration, seasonality was identified in hourly blocks, so an LSTM with 24 hour time steps was used to capture this pattern. As our paper sets out to investigate multiple forecasting techniques, the seasonality will need to be incorporated through different methods and approaches.

As the complexity of electricity demand has evolved so has the research into techniques for its forecasting. Earlier models explored Box and Jenkins time series modelling (Hagan and Behr 1987) or what is commonly known today as ARIMA. They found that a seasonal ARIMA model was well suited to electricity load forecasting where the forecast model was a linear combination of previous days electricity load with 24 hour (daily) and 168 hour (weekly) autoregressive lag included in their model. It was noted that including a transfer function to model the non linear relationship between temperature and electricity load should improve the forecasting error. A similar autoregressive approach in VuD.H.2017Sedf using NSW daily electricity demand data looked at grouping days of the week with similar electricity demand profiles together (weekdays and

weekends) in the modelling to improve the robustness of the forecast and to vary its coefficients with time to better capture the relationship to more recent data points. Leveraging the predictive performance of lagged demand discussed in the autoregressive literature above, our paper will utilise lagged demand as a predictive variable in our models.

As previously discussed, AEMO's half hourly model is a linear regression technique that uses the LASSO regularisation algorithm to produce a forecast which aims to 'describe the relationship between underlying demand and key explanatory' (Australian Energy Market Operator 2023b). In this respect it is more of an explanatory model in contrast to the predictive ARIMA methods above that have no explanatory power. Multiple regression techniques have provided a simple, yet powerful approach to forecast short term electricity demand. In (Charlton and Singleton 2014) a study was performed on 20 different locations, with 24 linear regression models generated to forecast electricity demand for each hour of the day at each location, which included variables such as temperature (and polynomial transformation), dummy variables for season and weekday/weekend with good performance of 0.70 R2. In the Australian context, S. Fan and Hyndman (2012) regression study on Victorian and South Australian electricity demand, produced a separate model (48) for each half-hourly period of the day, and included variables such as temperature (current and lagged, 24hr min and max), previous demand observations (24hr min, max and average) and calendar variables to forecast current electricity demand. A stepwise variable selection was carried out from a wider set of variables with a process of elimination and analysis of the effect on mean absolute percentage error (MAPE) used as the selection criterion. In "Understanding Intraday Electricity Markets: Variable Selection and Very Short-Term Price Forecasting Using LASSO" (2019), a linear regression forecasting model was used to predict 30min electricity pricing utilising the LASSO algorithm for its variable selection. This allowed an almost 'unlimited' number of variables to be used in the forecasting. In keeping with the approach by AEMO, we look to also implement a LASSO regularisation algorithm but our approach will investigate additional transformations and meteorological variables. While there was originally more focus on regression modelling of short term electricity demand, the research has shifted to techniques that can model the complex nonlinear relationship of electricity demand to temperature and seasonality. One of the most recent papers in the literature, Bashari and Rahimi-Kian (2020), forecasts 24 hour ahead electricity load for the city of Toronto by combining a LSTM network using historical electrical demand with a deep Feedforward Neural Network (FNN) using forecasted meteorological inputs (temperature). The combination of the two artificial neural networks outperformed each individual network as measured by Root Mean Square Error (RMSE). Another approach to model the non-linear relationships is forecasting with a Support Vector Regression (SVR) which uses a kernel to convert non-linear data to a high-dimensional space and then separates the data with a hyperplane. Jiang, Li, Lu, et al. (2020) utilised SVM algorithms with 30min time series electricity demand data to generate a forecast for both NSW and Singapore electricity demand, performing better than their benchmarking models. In Li et al. (2021), a least squares support vector machine was used to model half hourly electricity demand for New South Wales, Queensland and South Australia which as a method is less computationally expensive than the standard SVM technique. For this model, the data was split between work days and weekends and a lagged time series was used with a radial basis function (RBF) kernel to model each half hour forecast. Our paper will look to implement an SVM model experimenting with different kernels and test the impact of temperature which was not included in the literature above.

Finally, Random Forest regression models have also been used as a machine learning technique for short term electricity demand forecasting. They are a type of ensemble model that utilises classification and regression trees to apply bagging to generate a series or decision/regression trees that each vote on a result. One of the benefits of random forest models is they provide a feature importance metric that can be utilised for feature selection in high dimensions of data. In Huang, Lu, and Xu (2016), particular attention was paid to removing outliers as they could have a 'significant effect on the importance of the features'. The paper's analysis found that for 1 hour forward electricity demand forecast, the random forest model's most important features were the previous 24 hours of actual demand, whether it was a workday, and the day of the week. A more recent paper (Dudek 2022) applied the random forest method to forecasting short term electricity demand but adding additional features for daily and weekly seasonality. Again, only calendar features were used in the modelling for both papers, with no temperature data utilised. While there has been the use of random forest techniques for short term electricity demand forecasting in the Australian context (G.-F. Fan et al. 2021), they have been used within combined hybrid machine learning models, so we believe our

research into applying a solely random forest regression model for short term forecasting is novel, which will also seek to use meteorological data.

Our paper seeks to build upon the success in the literature above by applying similar machine learning methods (with additional meteorological data) to generate a comparison of the techniques to identify the best performing model specific to the NSW grid. In Z. Wang et al. (2021), it is apparent that the best model approach and subset of features are heavily dependent on characteristics (including geographic and demographic) of that area's grid, so while these machine learning methods above have been heavily researched, we seek to understand which one of them can outperform the AEMO demand forecast specifically for the state of NSW grids.

5 Material and Methods

5.1 Software

R and Python of course are great software for Data Science. Sometimes, you might want to use `bash` utilities such as `awk` or `sed`.

Of course, to ensure reproducibility, you should use something like `Git` and RMarkdown (or a Jupyter Notebook). Do **not** use Word!

5.2 Description of the Data

The data sets used in the analysis contained historical data from the state of NSW inclusive of the period 01-01-2010 to 31-01-2021 comprising historical electricity demand time series and meteorological observations over the period. Each data set has been stored in our public Github (Data Science Project: Group-L 2023). The data includes the following data extracts;

Total Electricity Demand

(totaldemand_nsw.csv 42.9mb)

Actual total electricity demand for the state of NSW in 5 minute increments measured in megawatts (MW). This data was sourced from the Energy Market Management System database of AEMO (Australian Energy Market Operator 2023a) made available by UNSW for the purposes of this research project.

Total Forecasted Electricity Demand

(forecastdemand_nsw.csv.zip.partaa 94.4mb, forecastdemand_nsw.csv.zip.partab 21mb)

Also known as the 'Half Hourly forecast', it is the forecasted electricity demand measured in MW for the state of NSW for a series of future 'half hourly' time series increments generated by AEMO per the 'Forecasting Approach Electricity Demand Forecasting Methodology' paper (Australian Energy Market Operator 2023b). This AEMO forecast data will be used as a benchmark to compare against the other forecasting models that will be investigated. This data was sourced from the Energy Market Management System database of AEMO (Australian Energy Market Operator 2023a) made available by UNSW for the purposes of this research project. The 1 hour forward accuracy of the forecast vs the actuals as measured by the Mean absolute percentage error (MAPE) was calculated at 3.30%.

Air Temperature

(temperature_nsw.csv 8mb)

Also known as 'dry bulb air temperature' (Bureau of Meteorology 2023c), air temperature in Celsius was taken from periodic (approximately 30min) observations from Bankstown Airport weather station and is used as a proxy to represent statewide temperature in NSW. It has been included in our analysis to include the well-known impact that temperature has to electricity demand (McCulloch and Ignatieva 2020). This data was sourced from the Australian Data Archive for Meteorology (Bureau of Meteorology 2023b), made available by UNSW for the purposes of this research project.

Solar Exposure

(solar_nsw.csv 416.4kb)

Solar exposure is the total amount of solar energy falling on a horizontal surface, measured in megajoules per square metre (MJ/m*m) (Bureau of Meteorology 2023b), with 'values highest in clear sky conditions during the summer, and lowest during winter or very cloudy days'. Solar exposure was included in this analysis to analyse the impact of solar photo voltaic electricity generation which acts as a net reduction to total electricity demand. The data used is based on daily solar observations derived from satellite data for the coordinates of Bankstown Airport weather station which will be a proxy for statewide solar exposure for NSW. This data was sourced from the Australian Data Archive for Meteorology (Bureau of Meteorology 2023b).

Rainfall

(rainfall_nsw.csv 744.5kb)

Rainfall in millimetres (mm) was taken from observations from Bankstown Airport weather station, with the data the representing daily amount of rainfall (Bureau of Meteorology 2023b). It will be used as a proxy for rainfall for the state of NSW. It has been included in the analysis to study the potential indoor/outdoor impact that rainfall may have on electricity demand. This data was sourced from the Australian Data Archive for Meteorology (Bureau of Meteorology 2023b).

5.3 Pre-processing Steps

R package In this project, data will be pre-processed via R. Data transformation will be completed via the tidyverse package (Wickham et al. 2019). Whereas date time related transformations will be completed via utilising the Lubridate package (Grolemund and Wickham 2011).

Data cleaning and preparation All data files used in this project will be in csv format. Data files will be read into R and processed to become fit for purpose via having the correct format and sufficient data integrity. Data will be processed to ensure consistency can be met, as well as having appropriate treatment for null values.

In terms of consistency, it would be to ensure each dataset has a consistent time gap between each measure when ordered by timestamp. This is important as the dataset will eventually be aggregated from by per minute to hourly, and inconsistent timestamp would result in hours having an unequal number of measures, where the number of measures per hour need to be equal in order for each hour to be unbiased. To achieve such consistency, each dataset used will be joined via left join to date-matrix of relevant timestamp intervals. This will effectively retain measures if they exist in the original data file, as well as surfacing timestamps where measures are null. Date-matrix data frames with intervals one, five, thirty minutes, as well as one day will be created.

Prior transformation takes place, all data files will be read into R, converted into dataframes. Each dataframe will be deduplicated to ensure all rows are unique. From counting the number of unique rows, the total demand dataset contains 39 duplicated rows therefore will be deduplicated within R. The R data frames generated will be filtered via relevant region columns to ensure data used is for NSW region only. Additionally, the intervals between each measure will be calculated via finding the datetime difference between current and previous measures, this column is used to determine the most suitable date matrix to join the dataset to.

Data transformation In this project, the five primary data sources utilised will be total energy demand, historical forecast, historical temperature, daily rainfall, and daily solar exposure. The timeframe of interest will be from 2010-01-01 to 2022-08-01. Data transformation will be explained separately by each dataset individually as per below:

Total Energy Demand The key columns required from the total energy demand dataset will be total energy demand and datetime. There are no extreme outliers in terms of total energy demand, therefore no filtering from the data is required. Given measures are mostly taken with five minute intervals, a five minute interval date-matrix will serve as the base table for left join.

Forecast Demand Forecasts generated by AEMO are nearly always on a half-hourly basis. Within each unique timestamp, AEMO will make multiple forecasts with the new forecast taking account of the previous, where forecasts are sequences via PERIODID (indexed from 1 ~ n, 1 being the first forecast, and n-th forecast being the final). Within the forecast demand dataset, all forecast values are within a reasonable range of having no outlier, therefore no filtering is required. In this project only the first and final forecast will be used, via transformation of the tables from long to wide format, removing the PERIODID and forecast measure columns, whilst adding two new columns for initial and final forecast. The final table post transformation will have three columns, datetime, initial forecast, and final forecast. The transformed table will then be left joined onto a date-matrix of thirty minutes.

Temperature The temperature dataset contains temperature measures with sporadic timestamps, usually between zero to thirty minutes, as well as larger gaps of up to three days. The temperature dataset contains extreme outliers such as -9999 degrees, whilst historically the lowest temperature ever recorded in NSW is -23 degrees celsius (Bureau of Meteorology 2023d). Therefore, the temperature data will filter out rows where temperatures are below -23 degrees celsius. The columns required from the temperature dataset will be datetime and temperature. The filtered temperature table with required column only will then be joined via left join to date-matrix with one minute intervals, due to sporadic timestamps mentioned above.

Daily Rainfall and Daily Solar Exposure For daily rainfall and solar datasets are recorded in daily format, with mostly 1 day intervals between each measure. The recorded measurements do not contain any extreme outliers therefore do not need to be filtered.

One thing to note with solar and rainfall datasets however, would be that the fields year, month, and day are in separate columns, hence to ensure the datasets can be joined to the date-matrix, the three fields will be combined into a single date field with yyyy-mm-dd format. Once the date field is created, the tables will then be joined via left join to date-matrix of one day interval.

5.4 Data Cleaning

To clean the data to ensure there are no null measurements for any timestamp, for any column that contains measurement, methods such as linear interpolation and last observation carried forward will be utilised (Y. Zhang and Thorburn 2022). Initially, simple imputation was considered due to simplicity, however because such a method is also prone to bias therefore will not be applied in this case (Emmanuel et al. 2021). Another idea was populating missing temperature via regression based interpolation methods (Xu et al. 2013). However, given the temperature data covers only one region, as well as missing timegaps being relatively small in almost all cases, linear interpolation will be sufficient due to simplicity. To summarise, missing total demand, temperature, daily rainfall, and daily solar exposure will be filled via linear interpolation, and missing AEMO forecast will be populated via last observation carried forward to prevent skewing from original forecast.

##Consolidation and Aggregation of final dataset Dataset Joins All data frames will then be joined into a single table in preparation for the modelling step. The cleaned dataset of total energy demand dataset will be used as base for join for other tables, given total energy demand will be considered the dependent variable for the model. Temperature and forecast dataset will be joined via datetime field. To join solar and rainfall dataset, a date field that does not include hour and minute will be derived from datetime column with yyyy-mm-dd format, solar exposure and rainfall datasets will then be joined to this field via relevant days.

Dataset Aggregation Given the objective of this project is to create a model capable of forecasting hourly demand, the combined data will be aggregated to hourly level. Given such, the total and forecast energy demand will be aggregated by summing the total/forecasted energy per hour. Temperature on the other hand will be averaged. Rainfall and solar will be untouched given the measures were initially at daily level.

Additional fields derived from existing data:

Year, Month, Day, Hour = Extracted from datetime field

Seasons = Derived from months, Spring = {9,10,11}, Summer = {12,1,2}, Autumn = {3,4,5}, and Winter = {6,7,8}

During Day = If the hour is between 8 am ~ 8 pm then 1, else 0.(this timeframe is selected due to having positive correlation to total energy demand, refer to Figure 9)

BusinessDay = If day between 1~5 then 1, else 0.

pub_holiday_flag = 1 if it is a public holiday, else 0. NSW Public Holiday table is retrieved via utilising the tsibble package in R (E. Wang, Cook, and Hyndman 2020)

Extreme temperatures of day and night fields will be derived via guidelines provided by the Australian government (Bureau of Meteorology 2023a). The fields will be in {1,0}.

Extreme Hot Day = Over 40 degrees and during day equals to 1 Hot Day = Over 35 degrees during day and during day equals to 1 Extreme Hot Night = Over 25 degrees and during day equals to 1 Hot Night= Over 20 degrees and during day equals to 1

Extreme Cold day = Under 10 degrees and during day equals to 0 Cold day = Under 15 degrees during day and during day equals to 0 Extreme Cold Night= Under 0 degrees and during day equals to 0 Cold Night= Under 5 degrees and during day equals to 0

##Assumption Assumption Based on the data used in this project, as well as various studies completed in the past, three main assumptions could be made with regards to behaviour of total energy demand upon interaction with various factors.

The first assumption is energy demand will increase during colder or hotter temperatures. Based on research completed in the past, it has been confirmed that winter and summer seasons have higher demand for energy during heating and cooling needs (Savic 2014). Additionally, according to research completed by Zhang et al. suggests energy demand tends to increase when temperature increases drastically (S. Zhang et al. 2022).

The second assumption would be in order for energy to be consumed, it would need to occur during the time of household resident activity. Past studies suggest that energy demand does indeed to be higher between 6 am to 9 pm, in contrast to 10 pm ~ 5 am (Savic 2014).

For the third assumption, given energy demand seems to increase during hotter and colder temperatures, this raises the possibility that if the range of global temperature were to increase over time, it would lead to the possibility of energy demand increasing over the years. According to a study completed by Jakob and Walland to investigate Australian extreme temperatures, it has been established that the extremely hot temperatures are becoming more frequent (Jakob and Walland 2016), which raises the question of potential increase in energy demand over the years.

5.5 Modelling Methods

6 Exploratory Data Analysis

Historical Trend of All Variables In figure 1, line charts are created at the most granular level of year-month-date-hour. However, at this level of detail, it is difficult to identify trends clearly due to the amount of noise presented, suggesting further aggregation is required in order to identify the trends clearly. Also to note that the final period recorded of 2022-08 is omitted from the charts, as the only datapoint from this particular month is incomplete at hour level, and would cause significant dropoff of trend.

Looking at the general trend over time for all variables from figure 1, the only variable that displays seasonality without significant variability is temperature per hour. Total energy demand per hour and daily solar exposure displayed seasonality albeit having a large amount of variability. Daily rainfall on the other hand did not display any seasonality, or any resonance towards changes of trend occurred in total energy demand.

Figure 2 displays a line chart of data aggregated to a month level. Aggregating data to monthly level displays a clear smoothing impact to the line charts.

Looking at Trend via individual years When the data is viewed at a monthly level via line chart in Figure 2, clear seasonality can be identified through having smoother lines in total energy demand, temperature, and solar exposure. Total rainfall viewed at monthly level still does not exhibit any seasonality. From an

assumption perspective, traces of energy demand being affected by temperature are starting to appear. Also with regards to assumption three, the trend chart of monthly total energy demand over time is starting to display a slow decline in total energy demand over the years.

When the monthly trend is viewed at individual year level as per Figure 3, it can be identified that the trend did not change significantly between 2010 and 2021, for energy demand, average temperature, and average solar exposure. The reason why 2022 is not used despite being the final year recorded is because the data used in this project does not contain all months of 2022. It can also be noted that total energy demand peaks when average monthly temperature is at its highest or lowest point. Additionally, solar exposure when magnified to monthly level surfaces its relationship with temperature clearly. Furthermore, rainfall's trend remained stochastic, exhibiting no relationship with other variables.

Total Energy Demand via Boxplot at Yearly, Monthly, and Hourly level Through using boxplot, it is possible to see the distribution of total energy demand across time. Figure 4 displays total energy demand is plotted against various time variables, specifically year, month, and hour. Each of these views will be discussed individually as per below.

Year At yearly view we can see the fluctuation of max and min total energy demand across time. The most key point taken from this view would be the slight decrease of average total energy consumption per hour across the years. This is an interesting point because it contradicts assumption 3, where an increase of total energy demand is expected due to an increase of extreme temperature occurrences. Thus, we can rule the assumption to be false, where yearly average hourly total energy demand is indeed decreasing.

Month At a monthly level, it can be identified that months of June, July, August, January, and February have higher average hourly total energy demand compared to the other months. This further verifies assumption 1 to be true, where total energy demand is expected to increase during hotter and colder months, as the months mentioned usually have the highest or lowest temperatures as per displayed in figure 2 and 3. Out of the previously highlighted months, June and July have the highest average total energy demand. This potentially shows that most people are more tolerant to hotter rather than colder temperatures, as June and July happen to have the coldest temperatures as per Figure 2 and 3. Furthermore, the boxplot of total energy by month suggests that months would be the best variable to use in terms of predictive modelling rather than seasons, as not necessarily all months within the season have higher total energy demand. For example, August is part of winter but has lower average total energy demand when compared to June and July.

Hour Looking at hourly boxplot of total energy demand, the key point identified would be total energy demand is highest between 8 am and 7 pm, and is lowest between 2 am to 6 am. This verified assumption 2, of total energy demand to be higher during hours of human activity. Hourly boxplot verifies that hour can be used as a variable in terms of modelling, as there is clearly a pattern exhibited by the chart particularly around the mean of total energy demand per hour.

Upon inspecting the densities of total energy demand by seasons and months via Figure 5, it can be further identified that the only group that has a visible different mean compared to the other seasons is winter. However upon looking at this at a monthly level, it can be identified that August does indeed have much wider distribution compared to June and July, as shown in figure 4 also. This further indicates that months should be used instead of seasons in predictive models in order to reduce noise.

Total energy demand by temperature, rainfall, and solar exposure

Scatter Plot of figure 6 has been created to display the relationship between total energy demand and temperature, rainfall, and solar exposure. At top level, all scatter plots created in figure 6 are exhibiting great spread, suggesting correlation to be more useful in terms of strength of relationship. Nevertheless, it is still possible to get some characteristics from these relationships, particularly for temperature. From the total energy demand against temperature scatterplot, it is possible to identify that total energy demand increases when temperature shifts either left or right from 20 degrees. Rainfall and Solar scatter plots on the other hand exhibited little trend, where rainfall scatterplot is very much just stochastic when compared to total energy demand.

Correlation Matrix of Total Energy Demand

Using the correlation matrix, it is possible to identify how strong are the relationships between each variable (Hadd 2021). When the data is inspected at correlation perspective via figure 7, from total energy demand perspective, the highest correlated variables are nearly always associated with cold temperature. For example, it is understandable where temperature does not hold a strong correlation with total energy demand, given in figure 6, where the scatter plot does not exhibit a linear trend; however, from the correlation matrix of figure 7, fields such as cold day, extreme cold day, and winter flag hold some of the highest correlation with total energy demand, indicating triggering those flags would result in higher total energy demand. Another field that holds a high correlation of 0.5 with total energy demand is during_day (flagging hours between 8 am ~ 8 pm), and this again strengthens assumption 2 of energy consumption being higher during human active hours. Furthermore, the negative correlation of -0.2 of datehour with total energy demand again contradicts assumption 3, hence we can settle with the statement that total energy demand is indeed gradually reducing over time.

In the correlation matrix of total energy demand against month flags, via figure 8, the highest correlation can be identified in July, as well as August, indicating higher energy demand during these two months. This is quite surprising as August had much less increase in terms of total energy demand compared to June and July as per Figure 4. This indicates that perhaps both July and August should be kept for creating the predictive model.

Upon inspecting the correlation matrix of total energy demand against hour flags in Figure 9, it can be identified that hour 3 to 5 have the highest negative correlation, whilst hour 17 ~ 9 have the highest positive correlation. The correlation matrix of the hour flag against total energy demand again demonstrates that there is a pattern between total energy demand and hour of day.

Overall, the correlation matrix from figure 7,8 and 9 shows us that in terms of predictive model, the fields that definitely should be used are extreme temperature flags, hour flags, and month flags.

Lagged Correlation of Total Energy Demand Finally, lagged correlation of total energy demand will be investigated. Lagged correlation looks at the correlation between time shifted time series data (Zagouras, Pedro, and Coimbra 2015). Essentially what this section is after is finding the best lag that can be used to predict future total energy demand. In this case, lags of 0 to 24 will be tested. The result as per figure 10, is that the more correlated lag would be lag of 1 and 24, which 24 will be ignored as it is essentially the same time given 24 hours. Therefore, total energy demand has the highest lagged correlation when shifted with a lag of 1.

7 Analysis and Results

8 Discussion

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

9 Conclusion

References

AGL Energy. 2020. "Spotlight on: Flexible Generation." https://www.agl.com.au/thehub/articles/2020/05/spotlight-on-flexible-generation, (accessed: 15.09.2023).

Alonso, Andres M., Francisco J. Nogales, and Carlos Ruiz. 2020. "A Single Scalable LSTM Model for Short-Term Forecasting of Massive Electricity Time Series." *Energies (Basel)* 13 (20): 5328.

Australian Energy Market Operator. 2017. "Maximum and Minimum Demand." https://www.aemo.com.au/energy-systems/electricity/national-electricity-market-nem/nem-forecasting-and-planning/forecasting-and-planning-data/nem-electricity-demand-forecasts/2017-electricity-forecasting-insights/summary-forecasts/maximum-and-minimum-demand, (accessed: 15.09.2023).

———. 2023a. "Energy Market Management System (EMMS)."

———. 2023b. "Forecasting Approach -Electricity Demand Forecasting Methodology." https://aemo.com.au/-/media/files/electricity/nem/planning_and_forecasting/nem_esoo/2023/forecasting-approach_electricity-demand-forecasting-methodology_final.pdf?la=en, (accessed: 17.09.2023).

Australian Renewable Energy Agency. 2021. "THE GENERATOR OPERATIONS SERIES: Report Two: Ramp Rates for Solar and Wind Generators on the NEM." https://arena.gov.au/assets/2021/08/the-generator-operations-series-report-two.pdf, (accessed: 15.09.2023).

Bashari, Masoud, and Ashkan Rahimi-Kian. 2020. "Forecasting Electric Load by Aggregating Meteorological and History-Based Deep Learning Modules." In *2020 IEEE Power & Energy Society General Meeting (PESGM)*, 1–5. https://doi.org/10.1109/PESGM41954.2020.9282124.

Bhattacharya, Mita, John N. Inekwe, and Perry Sadorsky. 2020. "Convergence of Energy Productivity in Australian States and Territories: Determinants and Forecasts." *Energy Economics* 85: 104538.

Bureau of Meteorology. 2023a. "About the Climate Extremes Analyses." http://www.bom.gov.au/climate/change/about/extremes.shtml, (accessed: 25.09.2023).

———. 2023b. "Climate Data Online." http://www.bom.gov.au/climate/data/, (accessed: 27.09.2023).

———. 2023c. "Observation of Air Temperature." http://www.bom.gov.au/climate/cdo/about/airtemp-measure.shtml, (accessed: 27.09.2023).

———. 2023d. "Rainfall and Temperature Records." http://www.bom.gov.au/climate/change/about/extremes.shtml, (accessed: 27.09.2023).

Charlton, Nathaniel, and Colin Singleton. 2014. "A Refined Parametric Model for Short Term Load Forecasting." *International Journal of Forecasting* 30 (2): 364–68.

Clements, A. E., A. S. Hurn, and Z. Li. 2016. "Forecasting Day-Ahead Electricity Load Using a Multiple Equation Time Series Approach." *European Journal of Operational Research* 251 (2): 522–30.

Data Science Project: Group-L. 2023. "Project-Group-l." https://github.com/llong-unsw/project-group-L, (accessed: 27.09.2023).

Department of Climate Change, Energy, the Environment and Water. 2021. "Australian Electricity Generation - Fuel Mix Calendar Year 2021." https://www.energy.gov.au/data/australian-electricity-generation-fuel-mix-calendar-year-2021, (accessed: 15.09.2023).

Dudek, Grzegorz. 2022. "A Comprehensive Study of Random Forest for Short-Term Load Forecasting." *Energies (Basel)* 15 (20): 7547.

Emmanuel, Tlamelo, Thabiso Maupong, Dimane Mpoeleng, Thabo Semong, Banyatsang Mphago, and Oteng Tabona. 2021. "A Survey on Missing Data in Machine Learning." *Journal of Big Data* 8 (1): 140–40.

Fan, Guo-Feng, Meng Yu, Song-Qiao Dong, Yi-Hsuan Yeh, and Wei-Chiang Hong. 2021. "Forecasting Short-Term Electricity Load Using Hybrid Support Vector Regression with Grey Catastrophe and Random Forest Modeling." *Utilities Policy* 73: 101294.

Fan, Shu, and R. J. Hyndman. 2012. "Short-Term Load Forecasting Based on a Semi-Parametric Additive Model." *IEEE Transactions on Power Systems* 27 (1): 134–41.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with Lubridate." *Journal of Statistical Software* 40 (3).

Hadd, Alexandria. 2021. *Understanding Correlation Matrices*. Quantitative Applications in the Social Sciences. Los Angeles, CA: SAGE Publications, Inc.

Hagan, Martin T., and Suzanne M. Behr. 1987. "The Time Series Approach to Short Term Load Forecasting." *IEEE Transactions on Power Systems* 2 (3): 785–91.

Huang, Nantian, Guobo Lu, and Dianguo Xu. 2016. "A Permutation Importance-Based Feature Selection Method for Short-Term Electricity Load Forecasting Using Random Forest." *Energies (Basel)* 9 (10):

767–67.
Jakob, Dorte, and David Walland. 2016. "Variability and Long-Term Change in Australian Temperature and Precipitation Extremes." *Weather and Climate Extremes* 14 (C): 36–55.
Jiang, Ping, Ranran Li, Ningning Liu, and Yuyang Gao. 2020. "A Novel Composite Electricity Demand Forecasting Framework by Data Processing and Optimized Support Vector Machine." *Applied Energy* 260: 114243.
Jiang, Ping, Ranran Li, Haiyan Lu, and Xiaobo Zhang. 2020. "Modeling of Electricity Demand Forecast for Power System." *Neural Computing & Applications* 32 (11): 6857–75.
Kpodar, Kangni, and Boya Liu. 2022. "The Distributional Implications of the Impact of Fuel Price Increases on Inflation." *Energy Economics* 108: 105909.
Li, Ranran, Xueli Chen, Tomas Balezentis, Dalia Streimikiene, and Zhiyong Niu. 2021. "Multi-Step Least Squares Support Vector Machine Modeling Approach for Forecasting Short-Term Electricity Demand with Application." *Neural Computing & Applications* 33 (1): 301–20.
McCulloch, James, and Katja Ignatieva. 2020. "Intra-Day Electricity Demand and Temperature." *The Energy Journal (Cambridge, Mass.)* 41 (3): 161.
Savic, Aleksandar ; Milosevic, Stevan ; Selakov. 2014. "Cold and Warm Air Temperature Spells During the Winter and Summer Seasons and Their Impact on Energy Consumption in Urban Areas." *Natural Hazards (Dordrecht)* 73 (2): 373–87.
"Understanding Intraday Electricity Markets: Variable Selection and Very Short-Term Price Forecasting Using LASSO." 2019. *International Journal of Forecasting* 35 (4): 1533–47.
Vu, D. H., K. M. Muttaqi, and A. P. Agalgaonkar. 2015. "A Variance Inflation Factor and Backward Elimination Based Robust Regression Model for Forecasting Monthly Electricity Demand Using Climatic Variables." *Applied Energy* 140: 385–94.
Wang, Earo, Dianne Cook, and Rob J. Hyndman. 2020. "A New Tidy Data Structure to Support Exploration and Modeling of Temporal Data." *Journal of Computational and Graphical Statistics* 29 (3): 466–78.
Wang, Zhe, Tianzhen Hong, Han Li, and Mary Ann Piette. 2021. "Predicting City-Scale Daily Electricity Consumption Using Data-Driven Models." *Advances in Applied Energy* 2 (C): 100025.
Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy McGowan, Romain Francois, Garrett Grolemund, et al. 2019. "Welcome to the Tidyverse." *Journal of Open Source Software* 4 (43): 1686.
Xu, Cheng-Dong, Jin-Feng Wang, Mao-Gui Hu, and Qing-Xiang Li. 2013. "Interpolation of Missing Temperature Data at Meteorological Stations Using p-BSHADE." *Journal of Climate* 26 (19): 7452–63.
Zagouras, Athanassios, Hugo T. C. Pedro, and Carlos F. M. Coimbra. 2015. "On the Role of Lagged Exogenous Variables and Spatio Temporal Correlations in Improving the Accuracy of Solar Forecasting Methods." *Renewable Energy* 78 (C): 203–18.
Zhang, Guoqiang, and Jifeng Guo. 2020. "A Novel Method for Hourly Electricity Demand Forecasting." *IEEE Transactions on Power Systems* 35 (2): 1351–63.
Zhang, Shaohui, Qinxin Guo, Russell Smyth, and Yao Yao. 2022. "Extreme Temperatures and Residential Electricity Consumption: Evidence from Chinese Households." *Energy Economics* 107: 105890.
Zhang, Yifan, and Peter J. Thorburn. 2022. "Handling Missing Data in Near Real-Time Environmental Monitoring: A System and a Review of Selected Methods." *Future Generation Computer Systems* 128: 63–72.

Appendix