CAPSTONE PROJECT BY TEAM GROUP K

FORECASTING ENERGY DEMAND IN NEW SOUTH WALES: AN ANALYSIS OF TEMPERATURE, REGIONAL REFERENCE PRICE, HOLIDAYS, AND TIME SERIES DATA, WITH AN EXAMINATION OF DEMAND IN RELATION TO POPULATION GROWTH

Abdelrhman Dameen (z5427841), Md Nezam Uddin (z5339862), Pam Moodley (z5366156), Van Hai Ho (z3071030)

School of Mathematics and Statistics
UNSW Sydney

October 2023

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

We would like to express our sincere gratitude to several individuals and organizations for supporting us throughout this project. First, we wish to express our sincere gratitude to our supervisors, Sarat Moka, Wei Tian, and Armin Chitizadeh for their enthusiasm, patience, insightful comments, helpful information, practical advice and unceasing ideas that have helped us tremendously at all times in our research and writing of this report. Their immense knowledge, profound experience and professional expertise in analysis Quality Control has enabled us to complete this research successfully. Without their support and guidance, this project would not have been possible. We could not have imagined having better supervisors in our study.

We are also grateful to the following Group K team members: Abdelrhman Dameen, Md Nezam Uddin, Pam Moodley, and Van Hai Ho for their consistent support and assistance.

Finally, last but by no means least; also to everyone in Endgame Economics for providing the dataset and for the great insights, time, and effort especially Oliver Nunn and Dr. Jack Simpson.

Thanks for all your encouragement!

06 October, 2023.

Abstract

Be for formal

This research report investigates the factors influencing energy demand in New South Wales, with a focus on variables like temperature, price, holidays, and time-series data. The study also examines the relationship between energy demand and population growth.Energy demand forecasting is essential for maintaining a sustainable energy infrastructure and for policy planning. With the challenges posed by climate change and urbanization, precise forecasting models are increasingly crucial for efficient energy resource management.Despite various models exploring energy demand, few studies comprehensively analyze how specific factors such as regional reference prices and holidays impact demand, and even fewer examine these in relation to population growth in New South Wales.How do temperature, prices, holidays, and population growth impact energy demand in New South Wales? Can machine learning models effectively forecast energy demand?A variety of machine learning algorithms were employed, including Linear Regression, Multi-Layer Perceptron, Random Forest, Facebook Prophet, and XGBoost. These models were assessed using metrics like MAPE, MAE, and RMSE.Machine learning models, particularly XGBoost, can offer nuanced and effective forecasts of energy demand. Time-series data, regional reference prices, holidays, and population trends are significant variables influencing energy demand.Among the models used, XGBoost showed the best performance with a 7% MAPE score.Factors like temperature, price of energy, and holidays are significant predictors.This research enhances our understanding of the multifaceted factors affecting energy demand in New South Wales. It demonstrates that machine learning algorithms, especially ensemble models like XGBoost, can be powerful tools in energy demand forecasting. By acknowledging the importance of regional factors such as population growth and holidays, this study contributes to more accurate and region-specific energy demand models.

Contents

Chapter 1

Introduction

The volatility of energy supply and demand poses a challenge for suppliers to enter and remain profitable in the market. Accurately predicting and efficiently supplying energy to the grid is critical for profitability. A key determinant of energy demand is weather; heating is necessary when temperatures drop, and air conditioning becomes essential as temperatures rise. Therefore, this analysis aims to examine the influence of weather on energy demand, considering the significant effects of global warming and erratic weather patterns.

However, other variables also come into play. As the International Atomic Energy Agency suggests, "The analysis should be conducted with relevant and consistent macroeconomic and microeconomic data, so that electricity demand projections can be more reliable and consistent with demographic, economic and industrial development projections" ((**iea-world-energy-outlook-2016?**)). This sentiment is further echoed by Emami Javanmard and Ghaderi who state, "The increase in population and economic growth of countries has led to a rise in energy consumption, which has created several challenges and problems for governments and nations" ((**Javanmard-et-al-2023-energy-demand-forecast?**)).

Correct these

Considering these challenges, the project will explore daily and seasonal variations, as well as the impact of holidays on energy demand. By incorporating these factors, we intend to highlight the benefits of using machine learning models for more efficient demand prediction, a point also emphasized by a report stating, "The demand for energy continues to grow as the world's population increases. And to ensure we meet these demands, utility and energy companies need reliable energy demand forecasting" ((**eia-2020-global-electricity-consumption-population?**)).

Specifically, we will uncover hidden temporal trends in demand, including daily and seasonal fluctuations, through the identification of energy consumption patterns. "According to the Global Energy Statistical website, energy consumption worldwide has increased by approximately 70% from 1990 to 2020" ((**perle-2023-predicting-power?**)). We will try to understand how variables like temperature impact energy demand. Furthermore, we will analyze the effect of holidays and special events on demand, aiding better planning efforts.

In New South Wales, the government anticipates exponential population growth, but this prediction may not sufficiently account for the intertwined relationship between energy resources and population capacity.

As populations grow, the demand for energy escalates, putting strain on existing resources. This can make energy sources scarcer and more difficult to extract, exemplified by the need to mine deeper for coal or explore complex environments for oil. The scarcity leads to declining marginal returns in energy extraction, pushing

the quest for new energy sources. These new sources, in turn, can expand the Earth's carrying capacity, enabling further population growth.

Therefore, the correlation between energy availability and population size could imply that if energy resources are nearing their peak production rates, New South Wales might also be approaching its maximum sustainable population. Hence, planning for the future should factor in these variables to create more accurate and sustainable growth forecasts. "In economies experiencing rapid residential electricity consumption and burgeoning energy-intensive activities, there is a notable link between economic growth and electricity use. Specifically, in less developed non-OECD countries, per capita electricity growth more than doubled from 2000 to 2017. This is in stark contrast to the nearly flat trend observed in more developed OECD countries"[5].

We hypothesize that growing population correlates with increasing energy demand. To confirm or refute this, we will perform an analysis that may also reveal other influential factors. Consequent to our analysis, policy recommendations will be provided to aid in energy policy formulation, including diversification of energy sources to meet demand. We aim to develop a machine learning model capable of predicting future energy demand with high accuracy, incorporating all the identified variables.

Electricity demand forecasting is an indispensable tool for managing the power grid and ensuring a reliable supply. It is a complex task, influenced by a multitude of factors. While traditional forecasting methods have their merits, there is growing interest in employing machine learning algorithms such as Linear Regression, Random Forest, and XG Boost for more accurate predictions.

For model performance evaluation, metrics like Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) will be used, given the time-dependent nature of the data. These metrics will serve as complementary tools for a comprehensive evaluation of our forecasting models.

Two common approaches to forecasting energy demand are top-down and bottom-up. The top-down approach focuses on macro factors like the economy, population growth, and weather, while the bottom-up approach examines energy consumption at the individual or company level. Both methods contribute to understanding future energy needs.

The subsequent sections of this report are organized as follows: Chapter 2 presents a literature review, establishing the relevance and importance of our study for policymakers in New South Wales and the energy sector. Chapter 3 elaborates on the methods, machine learning algorithms, and evaluation benchmarks. Chapter 4 delves into the data, exploring descriptive statistics and outlier analysis. Chapter 5 prepares the data for the main model and explores various visual plots. Chapter 6 examines the relationship between total demand and the estimated population of New South Wales. Chapter 7 analyzes and compares the results to other metrics, while Chapter 8 discusses these results. Finally, Chapter 9 concludes the report, offering recommendations and addressing further issues.

Chapter 2

Literature Review

When looking at the previous studies, papers and approaches for electricity demand forecast wwe discovered the use of different methods manual and automated, techniques in a way of forecasting energy demand, some of them addressed more ML methods. Other studies examined the broad applications of ML in energy system. Our unique focus on Demand in NSW electrical demand prediction offers deeper understanding, to the point of challenging conventional wisdom about things like the connection between pricing and temperature and the relationship between population increase and forecast issues. It is important to note, however, that we discuss deeper ways in which grid search is used to find the optimal model, in addition to pure applications of some benchmark measures to analyse the performance of our primary model.

needs citation

In his paper, Graham Zabel examines the sum-of-energies concept in relation to population expansion, highlighting its broad-reaching ramifications. Zabel acknowledges a number of aspects, including population control and natural catastrophes, but contends that energy resources have an indirect influence on these concerns. Instead than going into depth on particular energy sources like nuclear or hydroelectricity, the study primarily focuses on the significance of fossil fuels in influencing world population trends. According to Zabel, determining Earth's carrying capacity requires a knowledge of the interactions between energy resources and population development.

The 2023 paper by Pelka, titled "Analysis and Forecasting of Monthly Electricity Demand Time Series Using Pattern-Based Statistical Methods," focuses on the vital duty of precisely predicting monthly electricity load (MEL). Given that energy must be produced in real-time, the article, which was published in Energies, emphasises the significance of exact projections for sustaining affordable and dependable power systems. Pelka investigates a range of forecasting models, including traditional techniques, neural networks, and deep learning. In order to make complicated interactions between variables easier to understand, the study introduces the use of pattern representation in statistical approaches. Additionally, it emphasises how crucial it is to comprehend stationary time series in order to make precise predictions. The paper analyses data from Poland and 35 European nations to highlight the difficulties in MEL time series data, including non-linear trends and seasonal changes.

Writing for the U.S. Energy Information Administration Ari Kahan stats that the world's power consumption is growing faster than its population. The emerging, non-OECD nations where the per capita power usage more than quadrupled between 2000 and 2017 are where this trend is most pronounced. In contrast, the use

of energy has mostly followed a flat trend in industrialised OECD nations. Improvements in lighting technology and other efficiency measures have helped to somewhat offset this increase in usage. Kahan points out that in less developed countries, using more electricity per person is directly tied to economic growth. However, this isn't necessarily true for big, industrialized nations. Using the United States as an example, where per capita energy usage differs significantly from state to state, he also notices notable within-country variations.

In their recent paper from 2023 titled "Energy demand forecasting in seven sectors by an optimization model based on machine learning algorithms," Majid Emami Javanmard and S.F. Ghaderi explored long-term forecasting energy demand in Iran up to 2040. They employed a range of machine learning algorithms, including ANN, AR, ARIMA, SARIMA, SARIMAX, and LSTM, and integrated them with mathematical programming. In especially for Iran, the study emphasises the urgent problem of rising energy demand as people and economies expand. To improve their integrated model, the authors carefully assess the forecast accuracy of each algorithm in each industry. By providing a multi-algorithmic approach that takes into consideration the intricacies and variances of many sectors, this study adds to the body of literature already in existence. Additionally, they assess the performance of their integrated model using five criteria for prediction accuracy, demonstrating that their approach yields predictions that are more accurate than those produced by independent machine learning methods.

Shereen Elsayed, Daniela Thyssens, Ahmed Rashed, Hadi Samer Jomaa, and Lars Schmidt-Thieme demonstrate in their recent 2021 paper titled "Do We Really Need Deep Learning Models for Time Series Forecasting?" that XGBoost outperforms deep learning models for time series forecasting. They also prove that time series forecasting has a long history of simple models, such as exponential smoothing and linear models, outperforming more complicated ones. The paper makes several bombastic statements that seem more like "scientific propaganda" than an honest endeavor.

1. The paper claims that GBRT predictions outperform "state-of-the-art" neural forecasting methods, most of which are five years old, casting doubt on the seriousness of all their experiments. In the authors' words: "Stronger transformer-based models, such as the temporal fusion transformer, rightfully surpass the boosted regression tree."
2. It is suspiciously convenient that all nine benchmark datasets considered in the experiments are high-frequency.

Recently the government of NSW introduced a energy efficient program it started in 2009 and will run until 2050. These incentives give businesses and residents who own a home as well as companies financial incentives to upgrade their appliances and power outlets to a more energy-efficient one.
The state construction codes have also been changed to ensure that all newly constructed homes have better and more effective energy ratings, which will lower the need for electricity even if the population is continuously growing.

Our analysis aims to establish a generalized model for forecasting energy demands in NSW, which has both long-term and short-term forecasting implications.

and it uses a grid_search model to look for the best model and It employs multi-objective models that consider various machine learning algorithms to improve forecasting accuracy. It aims to provide energy suppliers and policymakers with valuable insights for better demand management, thus ensuring a stable and reliable energy supply for the NSW population.

Chapter 3

Material and Methods

3.1 Software

Python and R/RSudio software are used to Analyse the data. Libraries and packages such as pandas, matplotlib, seaborn for Python and ggplot2, dplyr, caret for R are required in this analysis. RMarkdown, knitr are also utilized for putting the analysis together.

Scikit-learn is a machine learning library in Python, widely used in this analysis. The algorithms we used for forecasting such as Linear Regression, Multi-Layer Perceptron, Random Forest and XGBoost are all available in scikit-learn library.

For project management, cloud storage, version control and code collaboration GitHub gave us pro level access as a student.

3.2 Description of the Data

We will use the provided data sets as our core data for our analysis, including:

- *totaldemand_nsw.csv*: Total Demand data.
- *temperature_nsw.csv*: Temperature data.

The data will need further analysis and cleaning, including the removal of invalid and outlier data before they will be used to generate the demand forecast.

3.2.1 Total Demand data

The Total Demand data provided in file *totaldemand_nsw.csv* contains energy demand in 5-minute intervals from January 1, 2010, to August 1, 2022, for New South Wales. The data is in a comma-delimited file format, with columns labeled Datetime, RegionId, and TotalDemand. The RegionId consists only of NSW1.

3.2.2 Temperature data

The temperature data provided in file *temperature_nsw.csv* is in 30-minute intervals from January 1, 2010, to August 1, 2022, for New South Wales. The data is provided in a comma-delimited file format, with headings DateTime, Location, and Temperature (in Celsius). The source of the temperature data is the Bankstown weather location.

3.2.3 NSW Public Holidays

We would like to understand if public holidays would impact the energy demand and what is the pattern. NSW public holiday data is publicly available from NSW Government Industrial Relations website.

The data source was manually captured from the different sources, and therefore were in different formats. One file was created for each year to ensure it was simpler to track which years were needed to be found and captured.

3.2.4 Aggregated Price and Demand Data

Energy price is also another factor affecting the energy demand. We use the aggregated price and demand data publicly available at Australian Energy Market Operator (AEMO) website (Operator ([no date])).

Aggregated Price and Demand data is available by month from 1998 to current month. For the purpose of this project, we need 156 data files for the months from 2010 to 2022 in order to integrate with total demand and temperature data sets. These data files are merged into a single file in the same format with the following headers:

- REGION: NSW Region.
- SETTLEMENTDATE: Settlement date and time for every 5 minutes.
- TOTALDEMAND: Total demand at the settlement date and time.
- RRP: Retail Price.
- PERIODTYPE: Period Type.

From this data set, only the settlement date and RRP was utilised.

3.2.5 Population Data

We hypothesised that as population is growing, the demand for energy is also increasing. We will perform an analysis to understand if our hyphothesis is true; or if it is not true, what other factors might have influence the energy demand. Population data that we use is publicly available at Australian Bureau of Statistics (Statistics ([no date])). This data is used in our analysis and is available in our repository: https://github.com/van-hai-ho/ZZSC9020_Project_Group_K/blob/main/data/NSW%20estimated%20population.xlsx.

3.2.6 Data set format

All the data sets discussed above need to be integrated before continuing. When merging these data sets, there is a mismatch in frequency since the demand is in 5-minute intervals and the temperature data is in 30-minute intervals. There are approximately 1.3 million rows of demand data and 247,646 rows of temperature data. When merging the temperature and demand data, the demand data is grouped by 30 minutes and the mean is utilised, and merged with the temperature intervals.

Since this project aims to address questions regarding future energy demand, historical data will be utilized, and the data sets provided are an excellent starting point. Additional data will be sourced to enhance these data sets.

As we hypothesised that the energy demand would increase when the population is increasing, we also use population data from ABS to prove our hypothesis. The population data used in this project is available from our repository: https://github.com/van-hai-ho/ZZSC9020_Project_Group_K/blob/main/data/NSW%20estimated%20population.xlsx.

Additional factors that could influence the demand is the energy price. To identify the correlations, data from AEMO is utilised, which provides the average price on demand every 30 minutes. The aggregated price and demand data used in this project is also available from our repository: https://github.com/van-hai-ho/ZZSC9020_Project_Group_K/tree/main/data/Aggregated%20price%20and%20demand%20data.

All datasets found where either excel sheets or comma delimited files, and these files where read in and analysed.

3.3 Pre-processing Steps

Part a and b of forecastdemand_nsw.csv.zip were unzip and then concatenated into a single file.

The data type on the columns which contain the date time are not date time, therefore the column have been cast to datetime for better analysis. A binary column is utilised to indicate if the particular day is a public holiday or not.

3.4 Data Cleansing

3.4.1 Energy Demand

The energy data contained the following columns,i.e. Datetime, RegionId, and TotalDemand. Only Datetime and TotalDemand data was utilised. We did ensure all the data in the dataset related to NSW, i.e. checking the distinct regionId, which was found to be "NSW1"

3.4.2 Temperature

Whilst looking at the minimum and maximum temperatures in the dataset, the minimum was -9999 which is an invalid temperatures. These rows were looked at and removed. Analysis of the data which where greater than -9999 but under 0 degrees celcius where kept since these where valid temperatures. The dataset has a Location column, and checking this column there is only one value of 94766, which relates to the weather station where the temperature was recorded. The Location column was excluded from analysis, since it did not add value.

3.4.3 Population

The source data for population has records from the year 1981. For this report, data from the year 2010 was utilised since the demand and temperature data only starts from the year 2010. The population data is in 3 month intervals, therefore data from 1-Mar-2010 is utilised when merging the demand with population data.

3.4.4 Public Holidays

Since the public holiday data were manually captured from different sources, the data files had different formats. Therefore each year was analyses and cleaned before being merged into one file which contained the date and the name of the public holiday. In this merged dataset, duplicates where found on the date field, since some dates had 2 different names, or 2 holidays where over-lapping each other, e.g. Easter Monday and Anzac Day, and New Years day had different descriptions. One row was kept for these kind of duplicates. The Name of the holiday was kept to ensure when merging the data to the demand, temp, rrp dataset it could be used when deciding to set the IsHoliday column to 1.

3.4.5 Energy Price

Only settlement date and RRP was utilised from the RRP source. The RRP column consisted of values of -1000 which at first glance didn't look valid. Reading through the source website there is RRP where it can be negative. An assumption has been made,i.e. -1000 is valid.

3.5 Assumptions

The 'temperature' in temperature dataset is given for Bankstown suburb in Sydney. In this analysis this temperature is assumed to be applied for NSW state.

3.6 Modelling Methods

XGBoost (eXtreme Gradient Boosting) is a more direct route to the minimum error, converging more quickly with fewer steps, and simplified calculations to improve speed and lower compute costs. It outperforms other algorithms like Random Forest, Multi-Layer Perceptron (MLP), and Linear Regression for several reasons.

Handling Non-Linearity and Interactions: XGBoost can capture complex relationships and interactions in the data, even when they are non-linear. This is particularly important in scenarios where the relationship between features and the target variable is not well-described by a linear model.

Ensemble Learning: XGBoost is an ensemble method that combines the predictions of multiple weak learners (usually decision trees) to create a strong learner. This can lead to more accurate and robust predictions compared to individual models like Random Forest or a single MLP.

Gradient Boosting: XGBoost uses gradient boosting, which builds trees sequentially. Each tree corrects the errors of the previous ones. This allows XGBoost to focus on the harder-to-predict cases and learn from its mistakes.

Regularization: XGBoost has built-in L1 (Lasso) and L2 (Ridge) regularization, which helps prevent overfitting by penalizing complex models. This is especially useful when dealing with high-dimensional data.

Handling Missing Data: XGBoost can handle missing data internally. It automatically learns how to treat missing values during the training process, reducing the need for imputation or data preprocessing.

Feature Importance: XGBoost provides a feature importance score, which helps identify the most influential features in the model. This can be useful for understanding which features are driving the predictions.

Efficiency and Speed: XGBoost is highly optimized for performance. It's designed to be memory efficient and can be parallelized, allowing it to handle large datasets and train models relatively quickly.

Tuning Options: XGBoost provides a wide range of hyperparameters that can be fine-tuned to improve performance. This includes parameters controlling tree depth, learning rate, and regularization.

Wider Applicability: While Random Forest is based on bagging and tends to work well for a variety of tasks, XGBoost's gradient boosting approach can be particularly effective in situations where there are a large number of features or where predictive accuracy is crucial.

3.7 Measures of forecast accuracy

To measure the accuracy, we take the difference between actual and the predicted value by the model, which is also known as forecast error. The lesser the forecast error the more accurate the model. There are several accuracy measures. Depending on the problem's nature and the model's implications, selected accuracy measures are chosen. In this analysis the following accuracy measures are considered:

3.7.1 Mean Squared Error (MSE)

MSE is the average of the squared difference between the target and the predicted value by the regression model. MSE is calculated by the following formula:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Since it squares the differences, it penalizes even a small error. The intuition of squaring the error is to make the large errors appear big. It is easy to calculate but sensitive to outliers.

3.7.2 Mean Absolute Error (MAE)

It is one of the simplest accuracy measures, the mean of absolute difference between target and the value predicted by the model.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

MAE is more robust to the outliers as it takes only the absolute value of the error. MAE does not penalize the error as extreme as MSE. So, when there are outliers in the data, MAE is preferable to use.

3.7.3 Root Mean Squared Error (RMSE)

It is simply the square root of the mean of squared errors and calculated as:

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

RMSE is used to compare forecasting errors of different models for a particular dataset and not between the datasets (Wikipedia, 2023). RMSE is always non-negative and a zero indicates a perfect fit. The lesser the RMSE the better the model fits the data. This accuracy measure is sensitive to outliers.

3.7.4 Mean Absolute Percentage Error (MAPE)

MAPE is a measure of prediction accuracy in a forecasting model (Wikipedia, 2023). MAPE is calculated by the following formula:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}|\frac{(actualvalue - forecastvalue)}{actualvalue}|$$

MAPE makes comparison of forecasting methods easier (and more useful) because working with percentage "standardizes" the errors. The time series' original units no longer matter. MAPE is affected by outliers.

Chapter 4

Exploratory Data Analysis

In energy demand forecasting there are several factors that can impact on demand and hence the forecasting. In this analysis for demand forecasting we consider several factors as features such as temperature, seasons, price of energy, holiday and several time series factors which could potentially impact energy demand. We tried to discover and investigate target and features relationship in this part of analysis.

4.1 Energy Demand Distribution

Figure 4.1 shows the density curve for the distribution of NSW energy demand.

Choose a better unit and mention it too

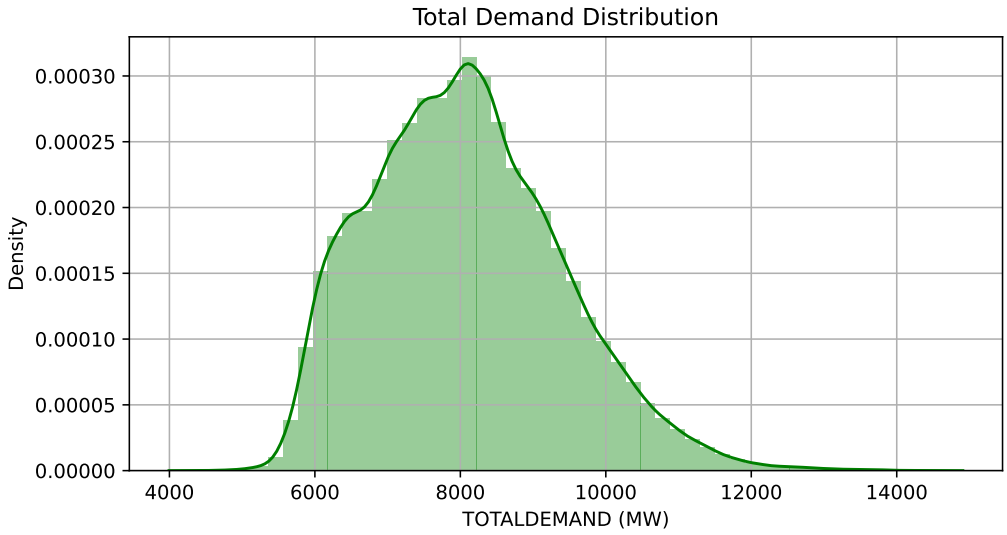


Figure 4.1: Total Demand Distribution

The distribution of energy demand is not symmetric, since an extended right tail is plotted, which exceeds 14500 MW of demand. This indicates there are occasional or a rare event when demand is significantly higher than the average demand. This density curve implies the energy provider and grid operators would need to be prepared for such rare high demand periods. To maintain a stable supply of energy the energy providers may need to have additional capacity in place.

4.2 Temperature Distribution

Figure 4.2 shows the density curve for the distribution of NSW Temperature.

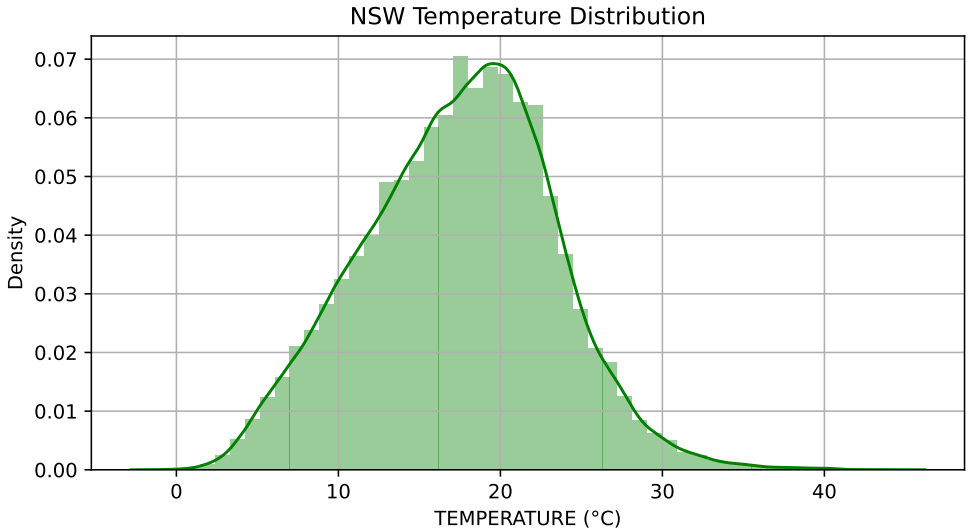


Figure 4.2: Temperature Distribution

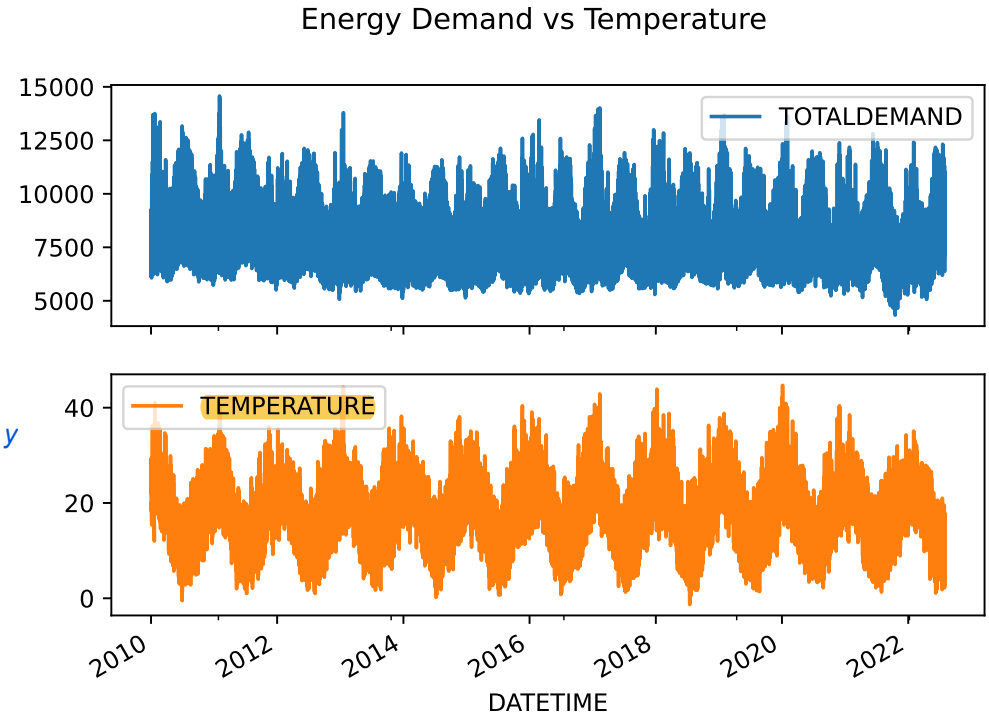
The distribution of temperature is not symmetric, and it is right skewed. The long tail to the right suggests majority of the data points (lower temperatures) are less than the mean temperature of 17.40°C, i.e. concentrated to the left and few number of data points (higher temperatures) are extending out to the right.

4.3 Relationship Between Energy Demand and Temperature

The variables which relate to weather, i.e. temperature, rainfall, solar exposure, wind speed and humidity may have a significant impact on energy demand. The inter-dependency between weather variables are complex, but temperature is the key influential climatic variable because it controls the atmospheric conditions. Therefore temperature has the most important impact on energy demand (Vu et al. (2014)).

Graphs of observed energy demand and temperature are plotted and also a Pearson correlation coefficient is measured to see if there is any linear relationship between temperature and energy demand.

```
## array([<Axes: xlabel='DATETIME'>, <Axes: xlabel='DATETIME'>], dtype=object)
```



Should be mentioned on y

Figure 4.3: Energy Demand vs Temperature

```
## Correlation coefficient:
##                TEMPERATURE  TOTALDEMAND
## TEMPERATURE       1.000000     0.114347
## TOTALDEMAND       0.114347     1.000000
```

Comparing the energy demand and temperature in Figure 4.3, it is observed that as temperature increases or decreases, the demand in energy consumption is increasing. The correlation coefficient between them is 0.114 which suggests a very weak linear relationship. But it is evident that temperature greatly affects demand, but the relationship is non-linear. A scatter plot confirms the non-linear relationship.

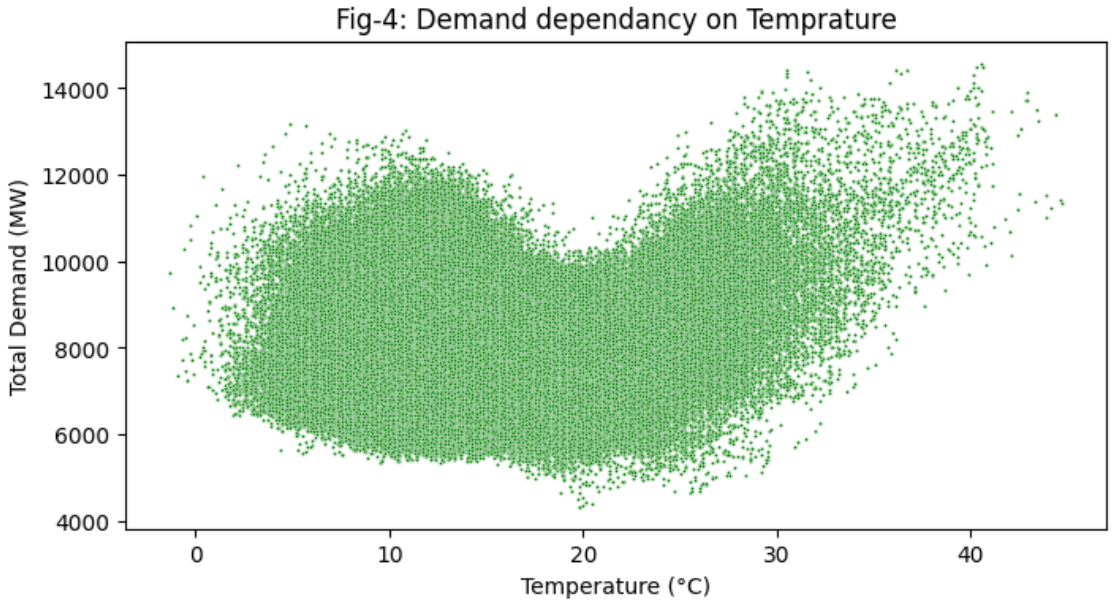


Figure 4.4: Energy Demand and Temperature Scatter Plot

The demand dependency on temperature in Figure **??** clearly exhibits a curve which indicates a non-linear relationship between demand and temperature.

4.4 Components of time series: Seasonality and Trend

Seasonality is a variation that occurs at specific regular intervals of less than a year (e.g. daily, weekly, monthly, or annually) and trend is the presence of a long term increase or decrease in the sequence of data (Auffarth (2021), Environment (2022)). This part of the analysis, some exploratory of the datetime identified a presence of seasonality trend with the NSW historical energy demand and temperature data.

Figure 4.5 shows the components of time series.

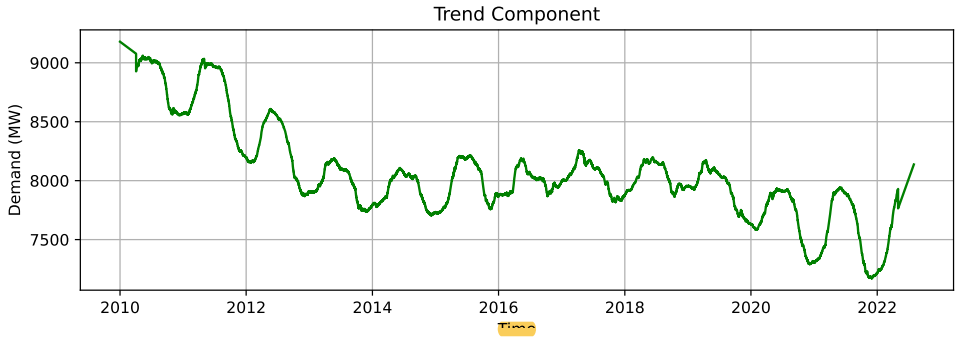


Figure 4.5: Energy Demand Yearly Trend

From the trend component in Figure 4.5, the demand for energy follows a downward trend since year 2010 which supports the data, i.e. the energy consumption in NSW has decreased by 2% over the past 10 years (NSW State of Environment

(2022)). Each year, there are regular spikes and dips in demand which indicate the presence of seasonality. For further investigation, the data is split into monthly and daily demand and can be represented in graphs.

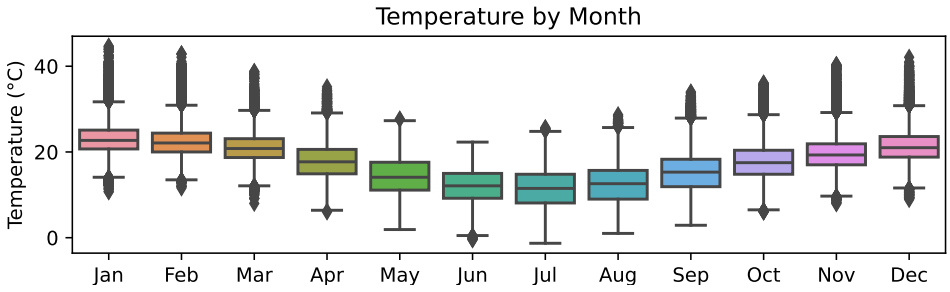


Figure 4.6: Energy Demand vs Temperature Monthly

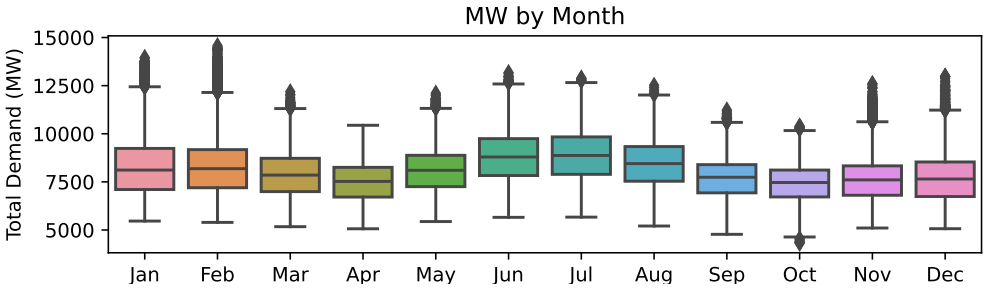


Figure 4.7: Energy Demand vs Temperature Monthly

In Australia, June, July and August are the coldest months during winter. Summer has three hottest months, i.e. December, January and February. Spring has three transition months September, October and November, and Autumn lasts for three months, i.e. March, April and May. The monthly energy demand in Figure 4.7 shows that during Summer and Winter the energy demand reaches its peak as people's usage of air-conditioning in summer and heating system in winter increases. It can be seen, during Spring and Autumn, the energy demand is lower, since the temperature remains at an average temperature and the need for heating or cooling homes decreases.

The amount of energy being used is affected by many factors, but mostly by temperature and time of the day.

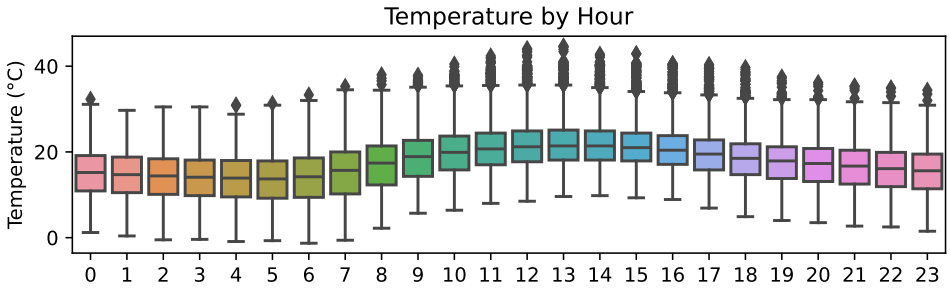


Figure 4.8: Energy Demand vs Temperature by Hour

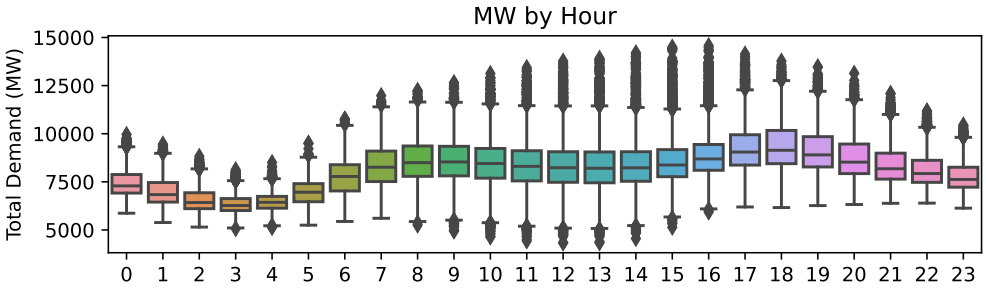


Figure 4.9: Energy Demand vs Temperature by Hour

From Figure 4.9, on an average day, it is observed the energy demand gradually increases through the daytime as temperature outside increases. The demand picks at around 6pm as people start getting home and use appliances when home. As the sun and temperature decreases, demand starts dropping off and reaches to its low between 2 and 3am in the morning because air-conditioning or heating equipment are not being utilised.

4.5 Relationship Between NSW Population Growth and Energy Demand

We assume that population has effect on energy demand, therefore, if population increases, energy demand will increase too. To test this assumption, we fit the regression line of energy demand against estimated resident population.

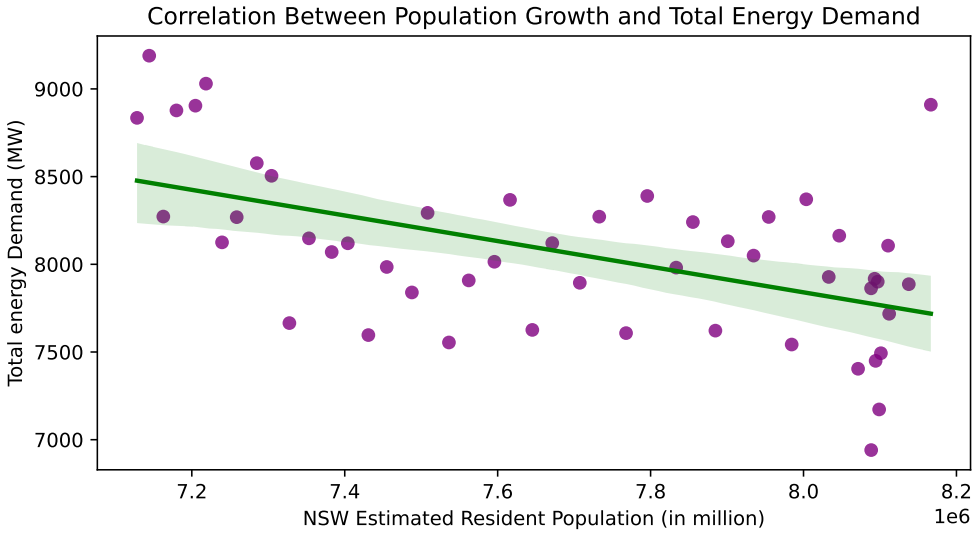


Figure 4.10: Correlation Between Population Growth and Total Energy Demand

As observed in Figure 4.10, the fitted regression line of energy demand has a downward trend, demand decreases as the population grows which also implies negative correlation. The calculated correlation is -0.53. For several reasons energy demand and population growth can exhibit negative correlation. Improved energy efficient technology and infrastructure, advanced appliances and industrial processes can significantly reduce the energy demand even if the population grows. Shifts to new energy sources- shifts from fossil fuel to more cleaner and efficient energy sources such as renewable energy can significantly reduce the energy demand.

4.6 Relationship Between Energy Demand and Energy Price

To investigate the relationship between energy demand and its price we fit a regression line of energy demand against energy price.

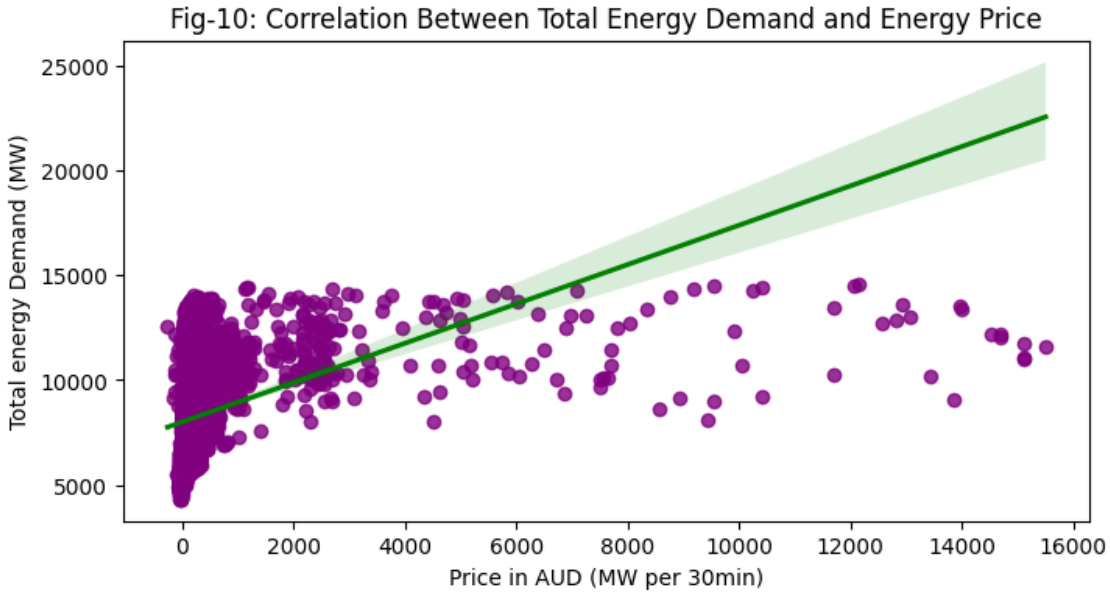


Figure 4.11: Correlation Between Total Energy Demand and Energy Price

The fitted regression line of energy demand in Figure ?? follows an uptrend which also implies positive correlation between energy demand and its price. The calculated correlation coefficient is 0.1345 implies a weak positive correlation. There are several reasons for which energy demand could increase even though price increases. For example, limited alternative to reduce the energy consumption in short term period, Seasonal and Weather effect, regardless of the price people are to consume more energy during winter and summer, income and economic growth leads to higher consumption of energy even if the price increases.

4.7 Relationship Between Energy Demand and Holiday

Most often people are away from home while on holidays which results in less consumption of energy. It is observed in Figure 4.12, energy demand tends to go higher when no holiday.

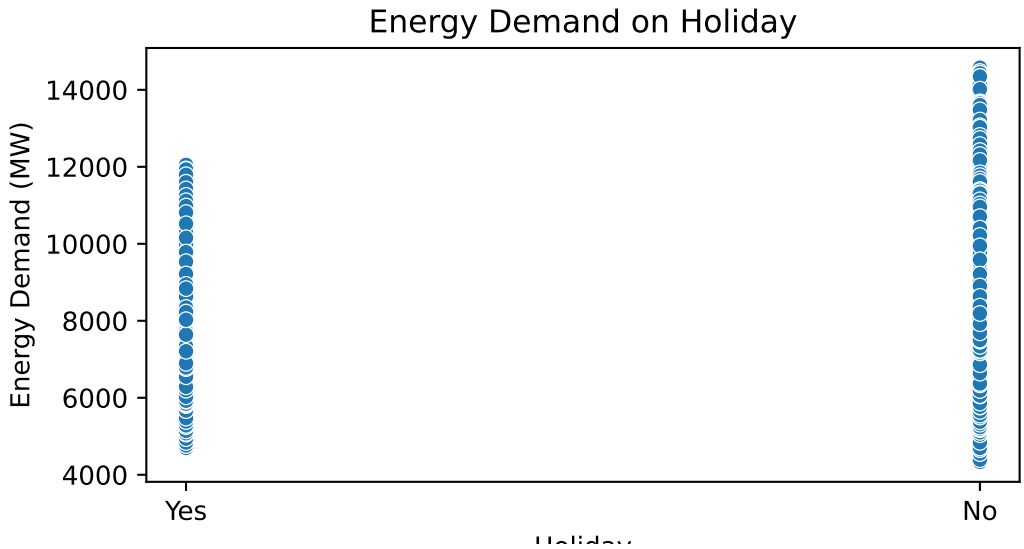


Figure 4.12: Energy Demand on Holiday

4.8 Demand Forecasting - Abdo deciding whether to add here.

Chapter 5

Analysis and Results

For the comparison of our model performance, we establish a benchmark using the given demand forecast data and actual demand data. We consider 30 minutes frequency over the periods of 48 that gives a 24-hour forecasting. The demand forecasting found to be highly accurate. Here below the MAPE, MAE and RMSE scores are given.

Table 5.1: Benchmark Evaluation Matrices

MAPE	MAE	RMSE
2.469%	204.027	280.450

5.1 Results from Initial Models

To forecast the energy demand, we initially consider three models: Linear Regression, Multi-layer Perceptron and Random Forest.

In the first experiment, target variable is energy demand and inputs are the temperature and time series features. The performances of different models are shown below:

Table 5.2: Evaluation Matrices - with Temperature and Time Series as Inputs

Model	Data	MAPE	MAE	RMSE
Linear Regression	train	10.606%	855.533	1080.704
	test	11.422%	883.416	1146.064
MLP	train	10.899%	887.541	1117.268
	test	13.721%	1007.229	1222.413
Random Forest	train	6.987%	573.809	763.936
	test	11.321%	833.064	1083.929

From Table 5.2, it is seen that Linear Regression model, with an MAPE error rate of 11.42% and RMSE of 1146.06 on test data performs better than Multi-Layer Perceptron. But Random Forest model outperforms both with an MAPE and RMSE score of 11.32% and 1083.93, respectively.

With Linear regression there is an assumption a linear relationship exists between the features and target variables. When plotting the Energy Demand against Temperature, it was found there was a non-linear relationship. Monthly, daily, weekly seasonality also exhibits non-linear relationship with energy demand. Since the relationship is non-linear, the linear model does not perform well with the complexity of the data.

Here Random Forest performed better than Linear Regression because:

- data are not normally distributed
- too many outliers in the data
- Target-Features relationship are non-linear
- Linear Regression struggle with complex feature interactions

Here Random Forest (RF) performs better than Multi-Layer Perceptron because:

- RF deals with outliers very well in the data
- less tuning for random forest
- MLP struggle with complex feature interactions
- Data are in a tabular format.

From the figure below, Figure 5.1, we can see that the performance of Random Forest prediction is average. It mostly fails to predict the demand when the demand is high or low, therefore, seasonality was not captured.

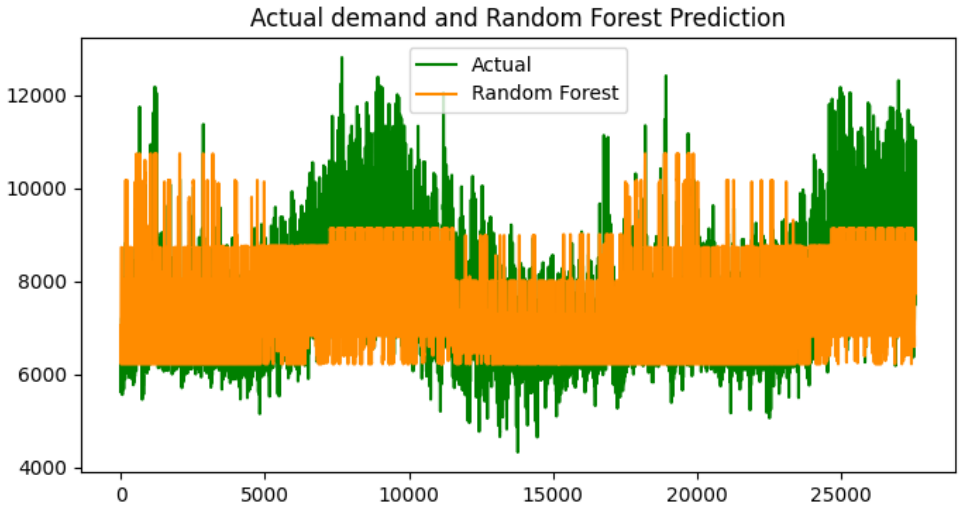


Figure 5.1: Actual Demand vs Random Forest Prediction

5.2 Facebook Prophet

As a comparison, Facebook Prophet is also considered. In the first model, to forecast energy demand, temperature, energy price, holiday and time series features are used as regressors. This model produces a Mean Absolute Percentage Error (MAPE) of 16.53. Again, a simple second model is built without any regressor. It is found out

that the second model with a forecasting error rate of 9.79 (MAPE) outperforms the first model.

Table 5.3: Forecasting error on test data set

Facebook Prophet	MAPE	MAE	RMSE
Model-1	16.527	1233.650	1520.480
Model-2	9.790	740.502	919.021

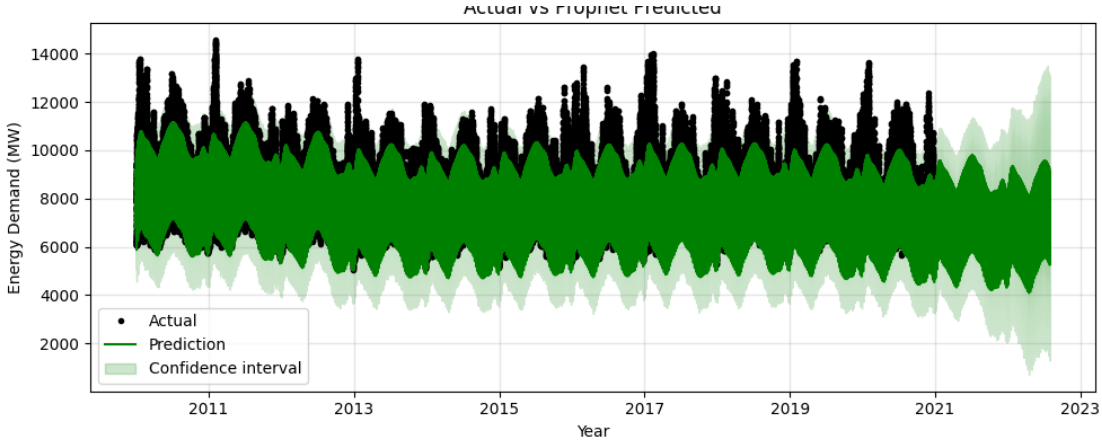


Figure 5.2: Actual Demand vs Prophet Prediction

5.3 XGBoost

In this final experiment, to forecast the energy demand by using XGBoost, we include three more factors, price of energy, holiday and one year lag as input. The reason for inclusions of one year of lag is that we are also doing a one-year demand forecasting for the period of Aug 2023 to July 2024. By taking the lag we are using historical demand data from one year ago and this would help us to capture the trend and pattern from the demand data which may persist into the future. The performances matrices are given below:

Table 5.4: Evaluation matrices- temperature, price, holiday and timeseries features are factors

Model	Data	MAPE	MAE	RMSE
XGBoost	train	5.87%	482.473	646.421
	test	7.03%	542.745	714.744

As compared to our initial models in the first experiment, XGBoost performed a lot better. It achieved a MAPE value of 7% and RMSE score of 714 on test dataset where for Random Forest they were 11% and 1083, respectively. The performance

scores between training and test datasets are not significantly different which indicates that this model has overcome the overfitting issue. XGBoost forecasting will be more reliable.

From the Figure: Actual demand and XGBoost Prediction, we see that it failed to predict some of the extreme demand such as demand which was over 1100MW and below 6000MW. But overall, it captured the most variation in the energy demand.

Figure: Actual_Demand_vs_XGBoost_Prediction

5.3.1 Future Forecast for one year

For next one year forecast we consider the period from Aug 2022 to July 2023. As we did not have future data, we considered the model performances on the training dataset. Our XGBoost model achieved an MAPE value of 5.76% and RMSE score of 629.

From the figure below, we can see that model was quite good, capturing the seasonal trend.

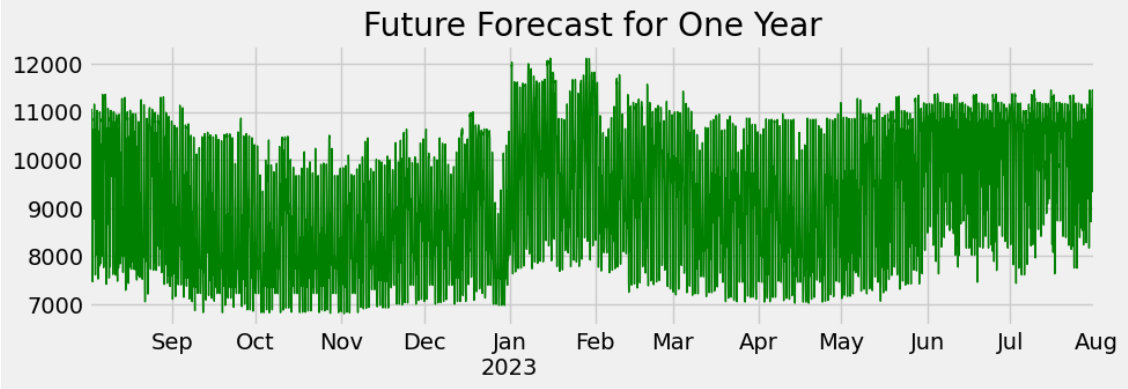


Figure 5.3: Future Forecast For One Year

Chapter 6

Discussion

Even though Random Forest beats Linear Regression and MLP but it's forecast error rate (11.32%) is quite high as compare to the benchmark error rate of 2.47%.

Facebook prophet beat Random Forest with a MAPE value of 9.8% and RMSE score of 740 but again it's error rate lot higher than benchmark error rate of 2.47%.

Moving forward, we agreed that there is scope to improve the accuracy of the prediction. Next, we considered another popular ensemble method XGBoost which is also based on decision trees and performs very well in capturing complex relationships and complicated patterns within the dataset. After trying several times with hyper parameter tuning our XGBoost model was able to reach an MAPE value of 7% and RMSE score of 715 which was the best among all the models we considered. So, we accepted this XGBoost model as our final model and then we applied this model for one-year future forecasting.

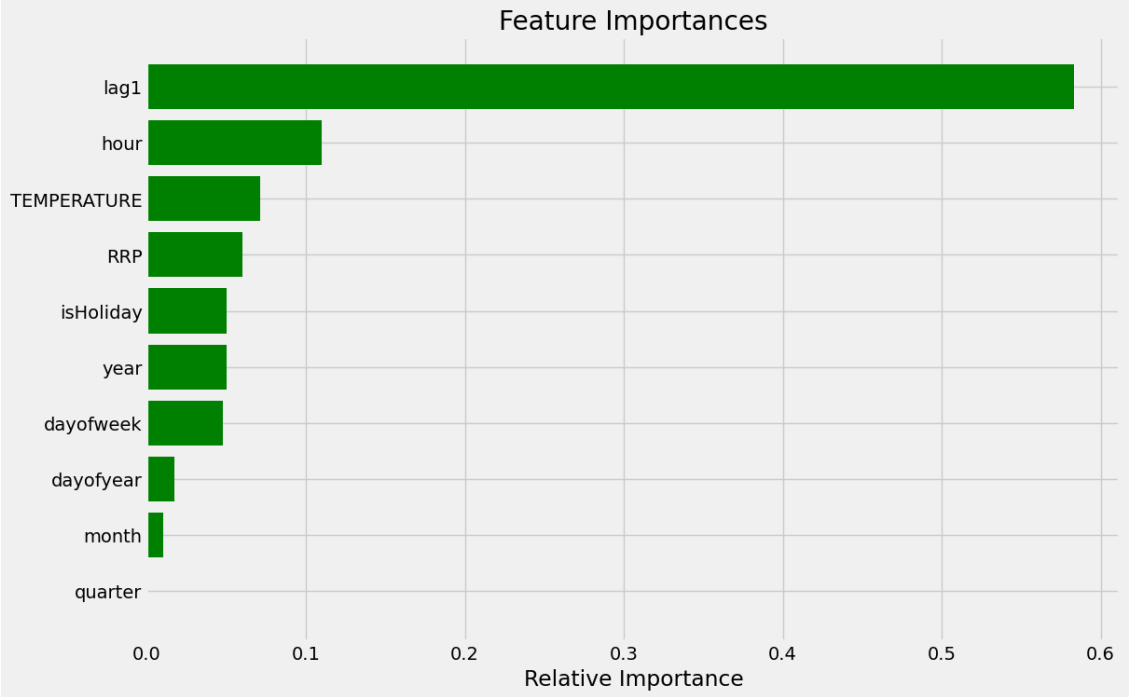


Figure 6.1: Feature Importance

The figure of Features Importance from final XGBoost model shows that in our energy demand forecast, one year lag and hourly lag were two most notable

features. Then Temperature, Price of energy and Public Holiday were the next key features which implies that there are relationships between energy demand and all these features. More importantly it proves one of our Project's hypotheses that Temperature, Price of energy demand and public holiday have impact on energy demand, and it is worth considering these factors when we do future forecasting.

Chapter 7

Conclusion and Further Issues

This paper has shown the impact of population, air temperature, seasons, price of energy, holidays and several time series factors which impact the energy demand. Population has a negative correlation with the energy demand, whilst the other features have an impact on the energy demand. From the other features, lag1, hour of the day, and temperature in the xgboost model have the highest importance, which means the previous years data, the hour of day and temperature had the most importance when predicting energy demand for the next year.

Recognizing the inverse relationship between population growth and energy demand observed in this study provides a valuable insight. The government of New South Wales should focus on sustainable practices and energy-efficient technologies to further reduce the environmental impact associated with increasing population densities, as highlighted "Population density in NSW has also risen. In June 2020, there were an average of 10.2 people per square kilometre – a 7.4% rise since 2015. Across Greater Sydney, the average density reached almost 480 people per square kilometre – 41 more than in 2015."(**NSWPopulation?**)

It is also important to note, its crucial to emphasize the need for future policies and strategies with the fact NSW government will need to manage the expectation of the energy suppliers, i.e. to keep the suppliers contributing to the energy grid with a promise of continued business prospects. The NSW government can take in consideration the effects of global warming, "Climate change is projected to increase temperatures in Sydney with maximum temperatures projected to increase by 0.7°C by 2030" (**NSWEnvironment?**) and with these increases in air temperature this would increase the energy demand.

More historical data would be required to improve the model. The other suggestions would be to include other features into the model which would contribute to the decrease of energy, and factors leading to the decrease in demand.

References

Auffarth, B. (2021) 'Machine learning for time-series with python'. Packt. Available at: https://subscription.packtpub.com/book/data/9781801819626/2/ch02lvl1sec11/identifying-trend-and-seasonality.

Environment, N.S. of (2022) 'Energy consumption'. Available at: https://www.soe.epa.nsw.gov.au/all-themes/human-settlement/energy-consumption#final-energy-consumption-status-and-trends.

Operator, A.E.M. ([no date]) 'Aggregated price and demand data'. Available at: https://aemo.com.au/en/energy-systems/electricity/national-electricity-market-nem/data-nem/aggregated-data.

Statistics, A.B. of ([no date]) 'National, state and territory population'. Available at: https://www.abs.gov.au/statistics/people/population/national-state-and-territory-population/latest-release.

Vu, D.H., Muttaqi, K.M. and Agalgaonkar, A.P. (2014) 'Assessing the influence of climatic variables on electricity demand'. in *2014 IEEE PES general meeting | conference & exposition.*, pp. 1–5. Available at: https://ieeexplore.ieee.org/document/6939377.

Elsayed, S, Thyssens, D, Rashed, A, Jomaa, HS & Schmidt-Thieme, L 2021, Do We Really Need Deep Learning Models for Time Series Forecasting?, arXiv.org, viewed 6 October 2023, https://arxiv.org/abs/2101.02118.

Zabel, G 2009, Peak People: The Interrelationship between Population Growth and Energy Resources - Resilience, Resilience.

Pełka, P 2023, 'Analysis and Forecasting of Monthly Electricity Demand Time Series Using Pattern-Based Statistical Methods', Energies, vol. 16, no. 2, p. 827.

Emami Javanmard, M & Ghaderi, SF 2023, 'Energy demand forecasting in seven sectors by an optimization model based on machine learning algorithms', Sustainable Cities and Society, vol. 95, p. 104623, viewed 11 September 2023, https://www.sciencedirect.com/science/article/pii/S2210670723002342

Environment, D of P and 2023, 'Sustainable building reforms offer long-term savings for households | NSW Government', www.nsw.gov.au, viewed 5 October 2023, https://www.nsw.gov.au/media-releases/sustainable-building-reforms#:~:text=The%20new%20standard%20cuts%20thermal.

(**NSWEnvironment?**) Fact sheet: Climate change in NSW - NSW environment and Heritage (no date) NSW Environment. Available at: https://www.environment.nsw.gov.au/-/media/OEH/Corporate-Site/Documents/Climate-change/climate-change-fact-sheet-160595.pdf (Accessed: 06 October 2023).

(**NSWPopulation?**) Population (no date) Population | NSW State of the Environment. Available at: https://www.soe.epa.nsw.gov.au/all-themes/drivers/population#:~:text=Population%20density%20in%20NSW%20has,41%20more%20than% (Accessed: 06 October 2023).

Appendix

7.1 Sourcing Public holidays

We did not take in consideration the local holidays since it changes depending on the council, and the current demand and temp data sets do not have local council data.

7.1.1 Source of data

- 2011: Nager.Date - Public Holidays in Australia 2011
- 2012: Nager.Date - Public Holidays in Australia 2012
- 2013-2015: NSW Public Holidays 2013-2015
- 2019-2020: NSW Public Holidays 2019-2020
- 2014-2024: Australian Public Holidays Dates Machine Readable Dataset

Put your own links like github and/or your files here too