UNSW SYDNEY

A Data Science Approach to Forecast Electricity Consumption in Australia Incorporating ENSO Data

ZZSC9020 – Data Science Project

Eugene Ho (z5497345) Data and Technical Specialist,

Majuwana Kariyawasam (z5398970) Data Specialist and AI Engineer,

Tariq Khan (z5414837) Data Specialist and ML Engineer,

Tom Woodley (z5450185) Group Leader.

05/10/2024

Table of Contents

Abstract

The accurate forecasting of electricity demand is of critical value to a wide range of stakeholders in Australia's energy market. There are numerous variables that influence energy demand, with meteorological conditions being a key factor. In turn, one of the main climate drivers in Australia is the El Niño-Southern Oscillation (ENSO). ENSO is a global climate phenomenon that cycles through three phases—El Niño, Neutral and La Niña— the former being associated with hotter and drier weather in eastern Australia, while the latter brings cooler and wetter conditions.

This study examines the ability of machine learning algorithms to take historical climate and energy data from New South Wales, and produce models of energy demand that are comparable to the industry standard forecasts from the Australian Energy Market Operator (AEMO). From here, it can be determined whether incorporating ENSO-related data as a predictor of energy demand can improve the performances of these models.

The techniques used were Linear Regression, Random Forest, XGBoost and Multilayer Perceptron Neural Network. With the RMSE of the AEMO's forecasted demand being 225MW, this was bettered by XGBoost which was the best performing model with a test RMSE of 195MW. This highlights the strength of this machine learning algorithm in the setting of forecasting energy demand. The key predictors of demand crucial to the success of these models were temperature, day of the week, population, and day of the year. Meanwhile, the inclusion of ENSO data was found to have no significant effect on the performance of any models.

1 Introduction

Forecasting energy demand in the Australian market is crucial to maintaining a balanced and reliable electricity supply. With fluctuations in weather conditions and seasonal patterns, it is essential for energy producers and operators to predict demand accurately, allowing them to adjust outputs and bidding strategies effectively. The Australian Energy Market Operator (AEMO) plays a central role in producing forecasts over a variety of timescales, which help energy producers, grid operators, and businesses manage their operations to prevent both shortages and surpluses (Australian Energy Market Operator, 2022). These forecasts form the backbone of energy supply planning, and their accuracy directly impacts grid stability, pricing, and operational efficiency.

In Australia, weather conditions are a primary factor influencing energy demand. Temperature variations in particular are significant drivers. For example, during periods of extreme heat, there is a marked increase in electricity consumption due to higher usage of cooling systems such as air conditioners. Conversely, colder spells can lead to a surge in heating demand which affects electricity consumption. Anticipating demand on the monthly, and longer timescales is important for legacy (fossil fuel) producers of electricity as their agility to react to changes in demand is determined by usage in the order of the hourly scale and also by securing adequate fuel to produce the energy.

AEMO closely monitors energy demand in response to unexpected events—referred to as "Shock Events"—which could be caused by sudden weather changes or infrastructure failures, further complicating the demand landscape (Australian Energy Market Operator, 2022). These shock events are commonly associated with extreme weather events on the climactic scale. One macro-driver of weather in the Pacific Ocean is the El Niño-Southern Oscillation (ENSO). ENSO is a climate phenomenon that affects weather patterns across Australia, with its two main phases—El Niño and La Niña—having a significant influence on rainfall, temperature, and extreme weather events. El Niño is typically associated with hotter and drier conditions in eastern Australia, while La Niña brings cooler, wetter conditions. These variations in climate have a direct effect on energy demand, especially in highly populated states like New South Wales (NSW), where electricity consumption is sensitive to seasonal weather fluctuations. As climate science advances, it is becoming increasingly possible to incorporate ENSO data into energy forecasting models. Given ENSO's role in driving large-scale weather changes, its inclusion in energy demand models is justified, particularly in regions like NSW, where climate variability can have profound effects on electricity consumption.

This project seeks to integrate data related to ENSO phases into more traditional weather and population data, and to feed this through common machine learning libraries in order

to produce models of energy demand. From here, investigations will be carried out to determine if these models can be competitive with AEMO's own projections. This focus is particularly relevant for legacy energy producers, who need to plan ahead for fuel procurement and production ramp-ups, as they cannot adjust output as quickly as renewable energy sources such as solar or wind. ENSO phases provide a valuable indicator of seasonal climate trends, and their inclusion could improve the ability of models to predict medium-term energy demand. In this project, the forecasting time scale is set to monthly intervals, as this aligns with the operational requirements of legacy energy producers while the granularity of the data remains at an hourly level to capture variations in demand at the dispatch cycle level (each dispatch cycle is half-hourly, so mean values are used to give hourly results).

Off-the-shelf machine learning models such as Linear Models (LMs), Extreme Gradient Boosting (XGBoost), Random Forest Regressors and Multilayer Perceptrons (MLPs) are used to assess whether the inclusion of ENSO phases enhances the predictive power of weather-based models when compared against the provided AEMO predictions. These machine learning models have proven effective in handling large and complex datasets, as well as identifying patterns in data and capturing non-linear relationships. Each of these abilities are critical when forecasting energy demand in a highly dynamic environment, such as an electricity market operating on robust infrastructure. Linear models, while not as accurate as more sophisticated machine learning models, are commonly used for predictions due to their simplicity and easy interpretability when engaging non-technical stakeholders. However, as the complexity of weather data and demand patterns increase, more advanced techniques such as XGBoost and Random Forest become essential due to their ability to capture intricate relationships between weather variables and demand, while not being limited by the assumptions of linearity and independence.

Incorporating ENSO phases as an additional predictor in these models represents a potential step forward in improving the accuracy of medium-term forecasts. In theory, the unique advantage of including ENSO lies in its ability to provide a longer-term view of climate patterns. Along with the recent advent of reliable ENSO forecasting, this may result in ENSO being useful as an explanatory variable in predictions. ENSO offers a broad seasonal perspective on climate which in turn allows for models that predict energy demand through secondary effects like temperature, wind and rainfall to improve accuracy. As an example: during an El Niño phase, energy producers in New South Wales might anticipate higher-than-average electricity demand due to prolonged periods of heat. ENSO also operates in a neutral phase or La Niña phase which drives lower temperatures and wetter conditions.

The choice of New South Wales as the focus for this study is strategic given it is located on the Pacific coast, therefore the assumption of it being affected by ENSO is most likely to be

valid. Furthermore, New South Wales is home to one half of the total population of Australia, and therefore represents a substantial portion of Australia's energy consumption. These two factors make it an ideal case study for examining the impact of ENSO on electricity demand.

Ultimately, this report aims to evaluate the effectiveness of incorporating a significant climate driver into a weather and population-based energy demand model to determine whether or not it can improve the accuracy of monthly forecasts. The results will be evaluated against the AEMO forecasts to determine if the models are competitive with the industry standard. The AEMO forecasts are likewise evaluated but calculating the root mean square error between its predictions and the actual demand over the time period investigated.

2 Literature Review

Numerous studies in recent years have explored the relationship between weather conditions and energy demand. It has been shown that temperature is a key driver of energy consumption at all levels of monitoring (International Energy Agency, 2023). Temperature appears to be the greatest driving factor in energy consumption, when controlling for population (Eskeland and Mideksa, 2010). This is especially true for regions with extreme climates such as NSW (Vu, Muttaqi and Agalgaonkar, 2015; Zhou et al., 2021). Research has also highlighted the importance of accounting for seasonal variations and extraordinary events in demand forecasting models (Albuquerque, Cajueiro and Rossi, 2022). The same research gives insight into the use of different modelling techniques for different time horizons. It is also of note that private (residential and commercial spaces) solar energy production is not captured by data published by AEMO due to the poor data surrounding private solar market penetration (Australian Photovoltaic Institute, 2024) and data capture for the weather-driver of solar energy production. Solar irradiance does not have sufficient granularity to determine production quantity using direct methods (Mfetoum et al., 2024) which makes it a prime candidate for machine learning and deep learning models. It would therefore be worth pursuing a machine learning model for solar cell market penetration using similar techniques to those presented here.

Many models rely on general linear models (GLMs) (Fan, MacGill and Sproul, 2017; Leung, 2022; Porteiro, Hernández-Callejo and Nesmachnow, 2022) which, while often accurate, struggle with higher-level interactions, seasonality and non-linear effects. GLMs are useful because of their simplicity and ease of communication to non-technical stakeholders. Additionally, they are not time or resource-intensive and can be trained and deployed easily (Nooruldeen et al., 2023). Many such models, such as ARIMA (Fattah et al., 2018) and its derivatives—ARIMAX and SARIMA (Kumar Dubey et al., 2021; Bilgili and Pinar, 2023)—are some of the most widely used because they have the advantage over other linear models of being able to deal with seasonal effects. The decision to use a linear model (scikit-learn's LinearRegression) as a basis for this report is therefore taken from the perspective that monthly forecasts do not experience the cyclic demand profile present in sub-daily and annual forecasts, which can be observed in the Exploratory Data Analysis. Linear regression models have been used to great effect at forecasting demand in the Sydney region under different parameters (Fan, Macgill and Sproul, 2015; Fan, MacGill and Sproul, 2017) but with the core assumptions in linear modelling of non-collinearity and linear relationships between explanatory and response variables. Given that the heatmap generated in the Exploratory Data Analysis shows spurious linearity between variables, it is necessary to investigate further using more sophisticated machine learning models.

Machine learning models such as Random Forest (Bedi and Toshniwal, 2019; Albuquerque, Cajueiro and Rossi, 2022; Vijendar Reddy et al., 2023), which is an ensemble method,

address non-linearities and seasonal influences effectively. Random Forest has an advantage over simpler models in that it is highly accurate and resistant to overfitting because it can handle non-linear model boundaries (AIML maintainers, no date). However, a challenge is communicating or visualising the complexity of the model generated, as well as diffiuclty easily extrapolating outside the range of given data (Thompson, 2019). This would be a distinct disadvantage when taking into account the effects of increased population, climate change and residential power generation (such as solar). More complex models such as Convolutional Neural Networks (Koprinska, Wu and Wang, 2018) and XGBoost (Extreme Gradient Boost)(Vijendar Reddy et al., 2023) have also been successful by handling seasonality, trends and non-linearities well. Similarly to Random Forest, XGBoost also struggles with extrapolation outside the range of the training set (Mavuduru, 2020) but is faster and often more accurate. A simpler neural network approach is also to use a multilayer perceptron (MLP) which could possibly combine the advantages of being both computationally cheap and able to handle seasonality and non-linearity (Scikit Learn Maintainers, 2024b, 2024a) whilst noting that, in practice, MLP are used to feed forward into other models for energy forecasting (Afzal et al., 2023). The application of a hybrid approach when employing an MLP belies an assumption that they are not sufficient for energy demand forecasting on their own. Both Random Forest and XGBoost will form part of the basis for this investigation and will each be representative of ensemble classification and gradient boosting algorithms. They will be applied along with linear regression and MLP models that have been tuned to fit the dataset. By applying all four modelling approaches, a picture of accuracy versus complexity can be determined.

None of the studies used to determine appropriate models integrate macro-seasonal factors such as ENSO, presenting a gap in current models. In the past two years, the ability to accurately predict when El Niño and La Niña will occur has dramatically improved (Liu et al., 2023), paving the way for a variety of statistics-based models to take advantage of this accurate forecasting. This work aims to expand on previous models by incorporating ENSO projections and population growth. A variety of off-the-shelf linear models and machine learning techniques (Random Forest, XGBoost and MLP) will be used to address this knowledge gap without considering hybrid models (Chreng, Lee and Tuy, 2022; Afzal et al., 2023).

An interesting further application, in addition to considering ARIMA, ARIMAX and SARIMA, is to use a hybrid model approach which blends two models or uses the output of one to feed into the other to produce accurate forecasts (Fan and Hyndman, 2010; Vu, Muttaqi and Agalgaonkar, 2015; Chreng, Lee and Tuy, 2022; Afzal et al., 2023). Additionally, the advanced linear-based technique: long short-term memory (LSTM) is also used in energy forecasts (Kumar Dubey et al., 2021; Mahjoub et al., 2022; Bilgili and Pinar, 2023; Qureshi, Arbab and Rehman, 2024) for its ability to overcome the extrapolation shortcomings of decision tree and gradient boosting algorithms. Both of these approaches are consistent with the state of

the art and could be used in future studies to build on the foundation of this project where boilerplate models were tuned and applied to a novel dataset.

3 Material and Methods

3.1 Software

The core platform of this project was Python. Python libraries for machine learning are intuitive and user-friendly which allows for easy collaboration. Typesetting and presentation in Jupyter Notebooks were equally user-friendly and thus used to present final results amongst team members and technical stakeholders as well as construct code that can be reproduced and altered as required. To present the final report, the Microsoft 365 suite of programs was selected due to its intraoperability and familiarity to non-technical stakeholders. The following libraries were used in this project:

- Pandas: Data manipulation and analysis
- NumPy: Mathematical functions
- Scikit-learn: Machine learning library
- XGBoost: Gradient boosting library
- Random Forest: Ensemble learning library
- PyTorch: Machine learning library
- Matplotlib: Data visualisation
- Seaborn: Data visualisation

The use of each of the above listed libraries was justified in the literature review where necessary.

For the non-technical requirements of the project:

- Discord was used for instant messaging due to its easy integration with mobile and computing platforms.
- Microsoft Teams was used for video conferencing and file sharing as well as meeting with supervisors.
- GitHub was used for version control and collaboration on code.
- Microsoft Planner was used for task management and tracking.
- Microsoft Word was used for the final report.

- Microsoft PowerPoint and OBS Studio were used for the final presentation and for recording the presentation.

3.2 Description of the Data

Historical electricity usage data of different regions were provided by course convenors. The datasets were processed and validated against the corresponding data collected from the Bureau of Meteorology (BoM) for the same period in the Sydney South-West region. Where the data was collected in text or Excel formats, they were converted to comma separated value (CSV) and stored on the GitHub repository. The datasets were cleaned and pre-processed before being used in the models. The final dataset contains only columns used for the machine learning models. The Exploratory Data Analysis section shows that different metrics for measuring the ENSO cycle were considered. Ultimately, the raw sea surface temperature and sea pressure differential data produced the best results.

The demand forecast file was too large to be stored on GitHub and therefore was converted to the parquet filetype. Parquet has an advantage over the CSV filetype for large files in that it uses a column orientation to decrease size dramatically but also retains the structure of traditional record-oriented files.

Population growth data from New South Wales produced by the Australian Bureau of Statistics (ABS) were also collected and used in the models. Additionally, population forecasts until 2070 were obtained from the ABS and used to forecast energy demand in the future.

ENSO cycle data: Southern Oscillation Index (SOI) and Sea Surface Temperature (SST) were sourced from BoM. BoM also publishes their own ENSO classifications, while SOI and SST were both processed to produce an additional set of ENSO classifications. The ENSO classifications were used in untuned models to determine their impact on model performance. It was determined (as is presented in the Exploratory Data Analysis) that the raw SOI and SST data produced the best results.

Since there were a great number of variables considered on the daily basis over the course of a decade, multiple files needed to be stitched together. The culmination of all the stitching was a file in the path: "../data/NSW/data_for_ml.csv" within the GitHub repository.

All data can be found in the GitHub repository under the data folder and data sources can be found in the appendices.

3.3 Pre-processing Steps

The provided weather data was validated for consistency with the historical weather data that includes temperature, precipitation and humidity from BoM, while superfluous data

were dropped. Weather observations were taken from as close as possible to the provided data's geographic location and temperature data was matched against the provided data to validate the efficacy of using data from a nearby weather observation point, given Bankstown Airport (the location of the provided dataset) does not provide all features we were interested in.

The data validation process mentioned in the above paragraph consisted of overlaying the provided temperature data with the collected data. Not only were the data observed to fall within the same range over the investigated domain but also trends and patterns were observed to be consistent.

All data files that came as .xls or .xlsx files were converted to .csv. The weather observation files were biennial and therefore had to be stitched together to form a complete data file for the time frame of interest.

The forecast demand data came as two .part files that required zipping together which was then too large for a GitHub repository. As a consequence, the file was converted from .csv to .parquet which solved the issue.

3.4 Data Cleaning

The velocity of data for this project was between half-hourly and daily and the forecasting timescale for this project was monthly, meaning that for the decade-long period over which the models were trained, there was sufficient data to drop all null rows. At most, one dataset had 374 missing rows from a total of 98280 which left sufficient data to then aggregate into the mean values of each column.

In addition to some missing values, some of the datasets had rows of explanation of their use in additional header rows which had to be removed. The header values were also changed in the ENSO and Population datasets to make them easier to interpret.

The data was examined in the exploratory data analysis either directly, by joining different datasets (which can be accessed through GitHub) or by using the "data_for_ml.csv" file. The code that was used to create the file was "join_data_for_ml.py" which can be found on GitHub and in the appendices.

3.5 Assumptions

Bankstown data is analogous to the Sydney South-West data and Sydney data can be extrapolated to the whole state. Both these assumptions are tested in this report. Firstly, by direct comparison in the exploratory phase (Appendix 2) and then by comparing the accuracy of the models in this report to the AEMO benchmark.

New South Wales is affected by ENSO in a similar way to all other Eastern states and therefore an investigation into New South Wales will suffice to draw conclusions for Queensland, Victoria, Tasmania and Australian Capital Territory.

3.6 Modelling Methods

The modelling methods used have been discussed in the literature review along with their justifications for inclusion. Those models are: Linear Regression, Random Forest, XGBoost, Multilayer Perceptron.

4 Exploratory Data Analysis

4.1 Temperature Distribution

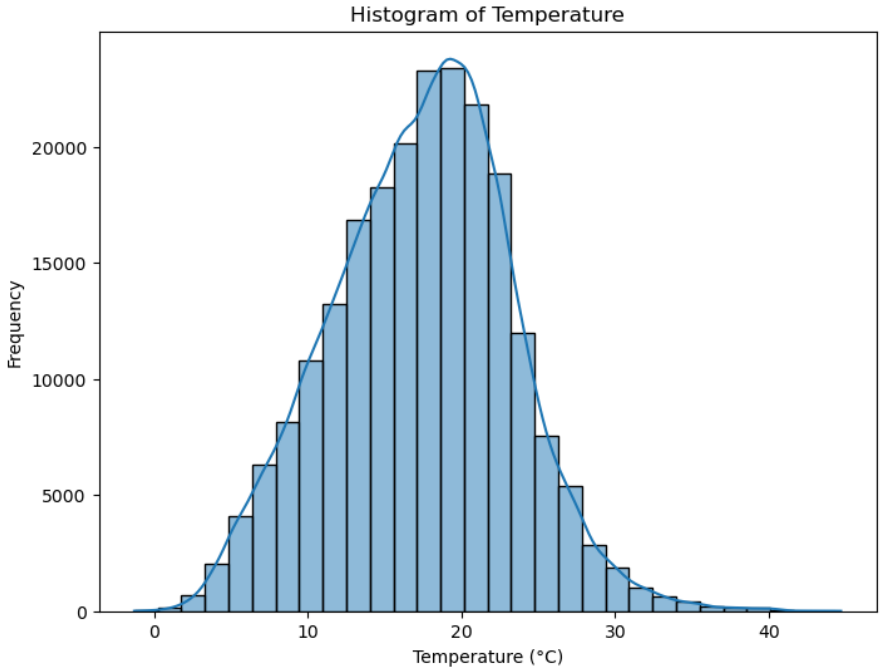


Figure 1. Histogram and distribution of temperature in the time frame.

Figure 1 illustrates the temperature distribution in NSW as well as its density curve. The distribution appears to be normal, centred around 20°C, with most temperatures ranging between 10°C and 30°C. There is a slight right skewness which indicates the presence of some higher temperature values that could be considered outliers. The density curve peaks near the average temperature.

4.2 Demand Distribution

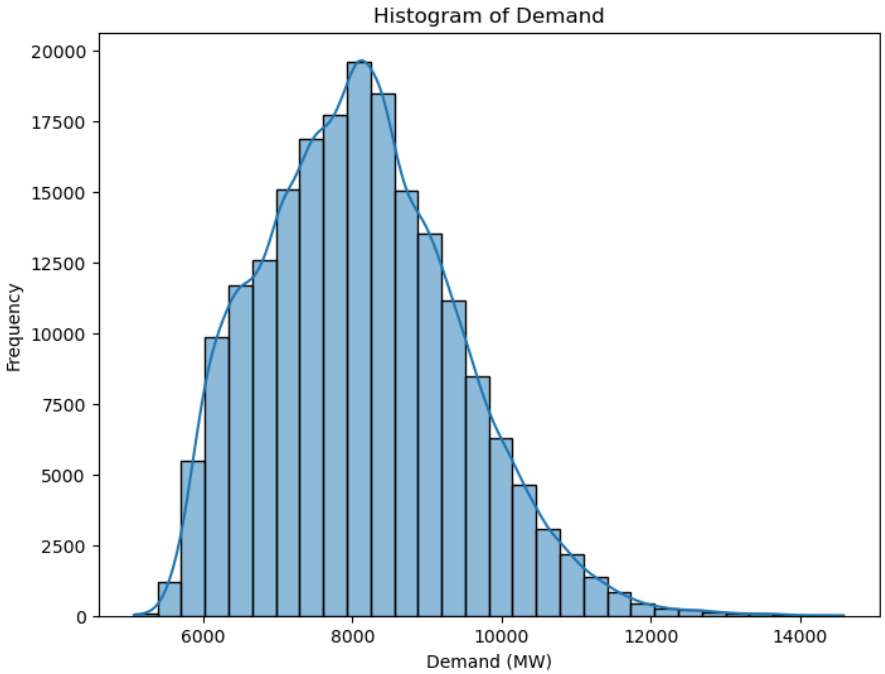


Figure 2. Histogram and distribution of total electricity demand in the time frame.

Figure 2 depicts the distribution of electricity demand in megawatts (MW). The distribution appears approximately normal with concentrated values around 8000 MW and most values are between 6000 MW and 10000 MW. The distribution shows a slight right skew, suggesting that there are some higher demand values, though they are less frequent. The density curve peaks at around the most frequent demand levels near 8000 MW.

4.3 Seasonality and Trend

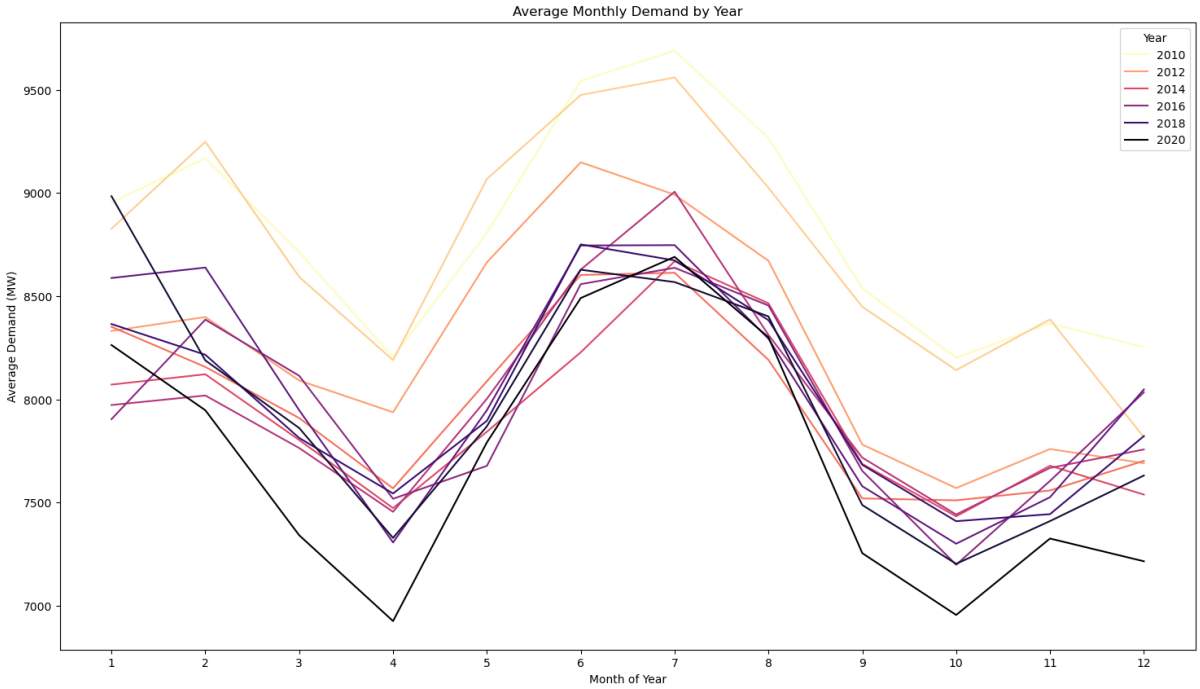


Figure 3. Line graph of average monthly demand for each year in the time frame.

The electricity demand depicted in Figure 3 seems to follow a cyclical pattern each year with higher demand during the summer months (around December, January and February) and winter months (June, July and August), likely due to higher heating and cooling needs during those months, while spring and autumn see lower demand. Therefore, it is likely that temperature markedly influences the demand of electricity.

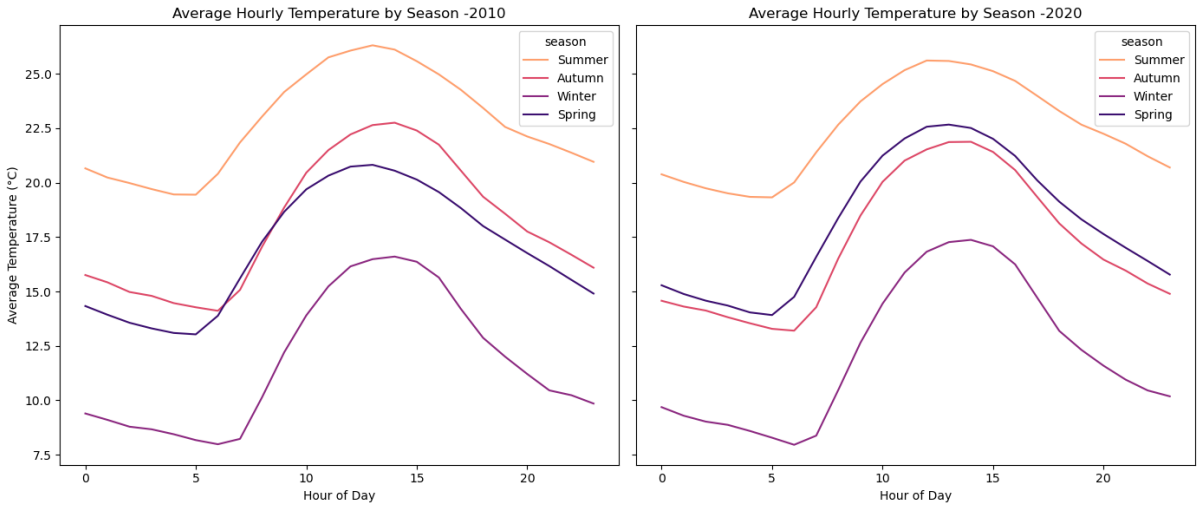


Figure 4. Average hourly temperature for 2010 (Left) and 2020 (Right).

Temperature peaks at around the same temperature and around the same time of the day, showing a consistent pattern throughout the years.

4.4 Relationship between Energy Demand and Temperature

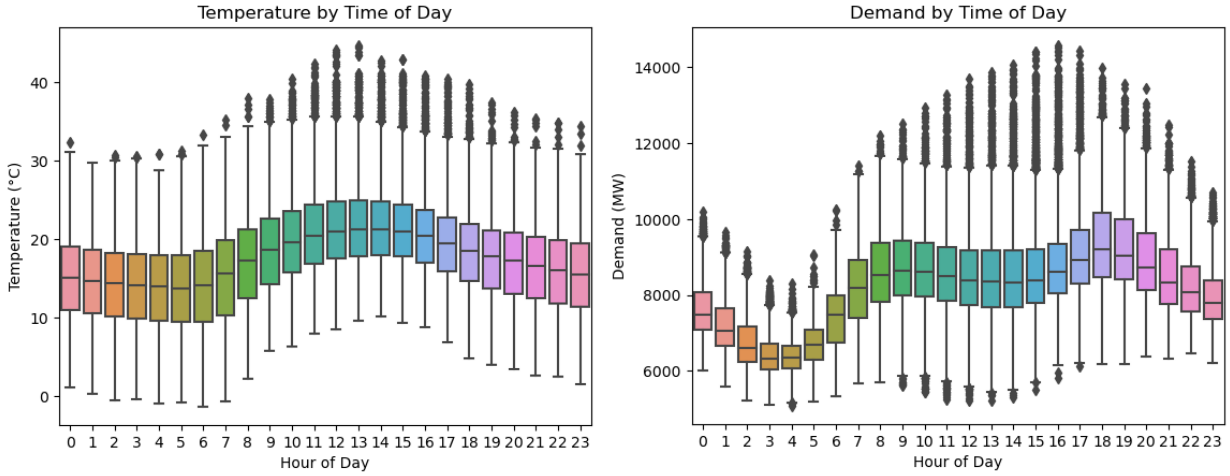


Figure 5. Comparison boxplots of temperature (Left) and demand (Right) for the time of day.

Further exploring the correlation between temperature and demand, Figure 5 shows boxplots of temperature during each hour of the day and the demand during each hour of the day. A very similar pattern to the above can be seen where demand for electricity increases as temperature increases. This further suggests that temperature and demand likely have some correlation.

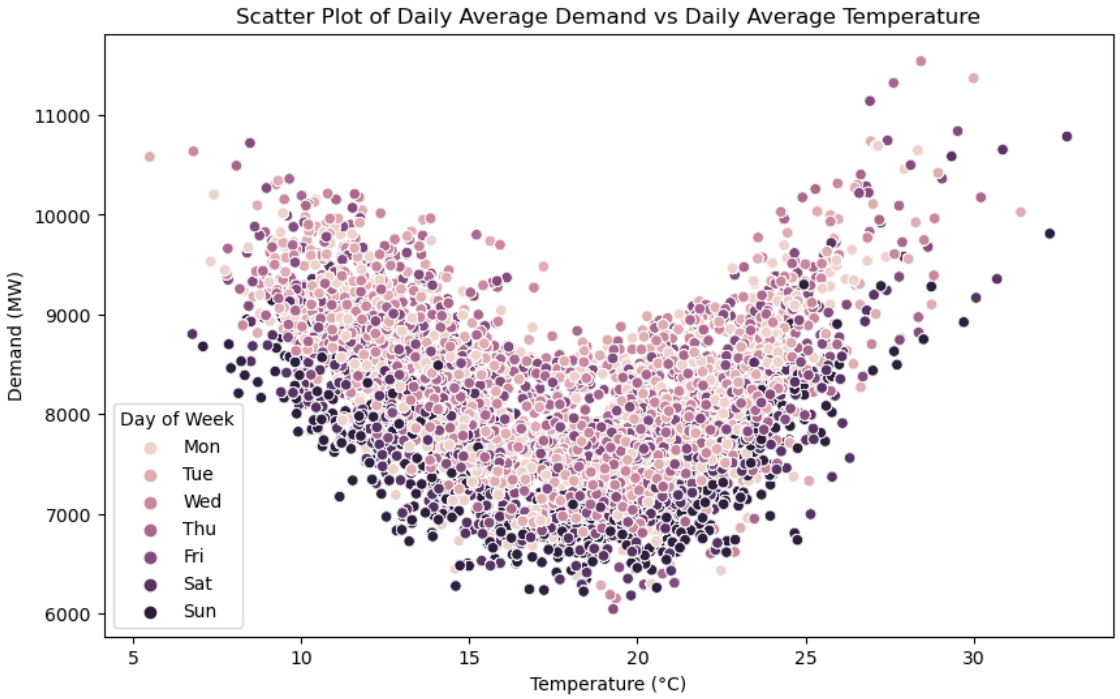


Figure 6. Scatterplot of demand against temperature, colour-coded for day of the week.

However, when we look at the demand and temperature scatterplot in Figure 6, in which we have separated each day of the week by colour, we can see that the relationship exhibits a U-shape with demand being higher at both low (5-10°C) and high temperatures (above 25°C), while it drops at moderate temperatures (15-20°C). This suggests that there is a non-linear relationship between Demand and Temperature, as both heating and cooling requirements drive higher electricity consumption.

It is also evident that the day of the week seem to have an influence on the demand level. Weekdays (Monday to Friday, lighter-coloured dots) generally have higher demand compared to weekends (darker-coloured dots). This pattern is consistent across the temperature range.

4.5 Population

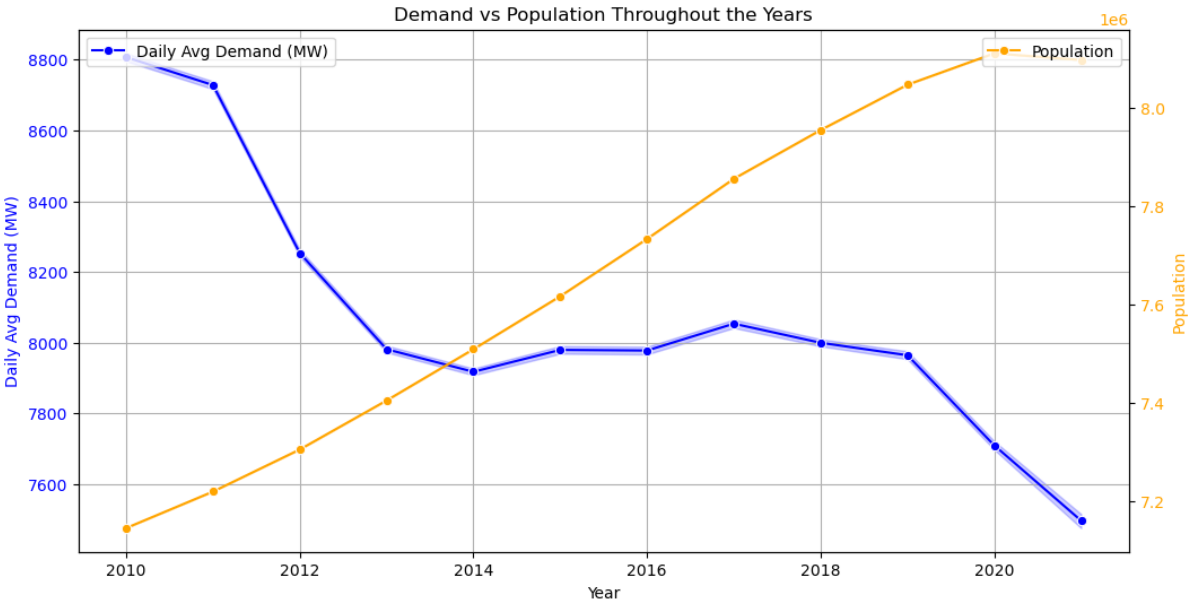


Figure 7. Line graphs of population (yellow) and demand (blue) showing opposite general trends.

Figure 7 shows us that despite the population growth, electricity demand has not followed the same upward trajectory. The divergence between population and demand trends indicates that external factors (beyond just population growth) are influencing electricity consumption, making it a point of interest for further analysis.

4.6 Relationship between electricity demand and ENSO

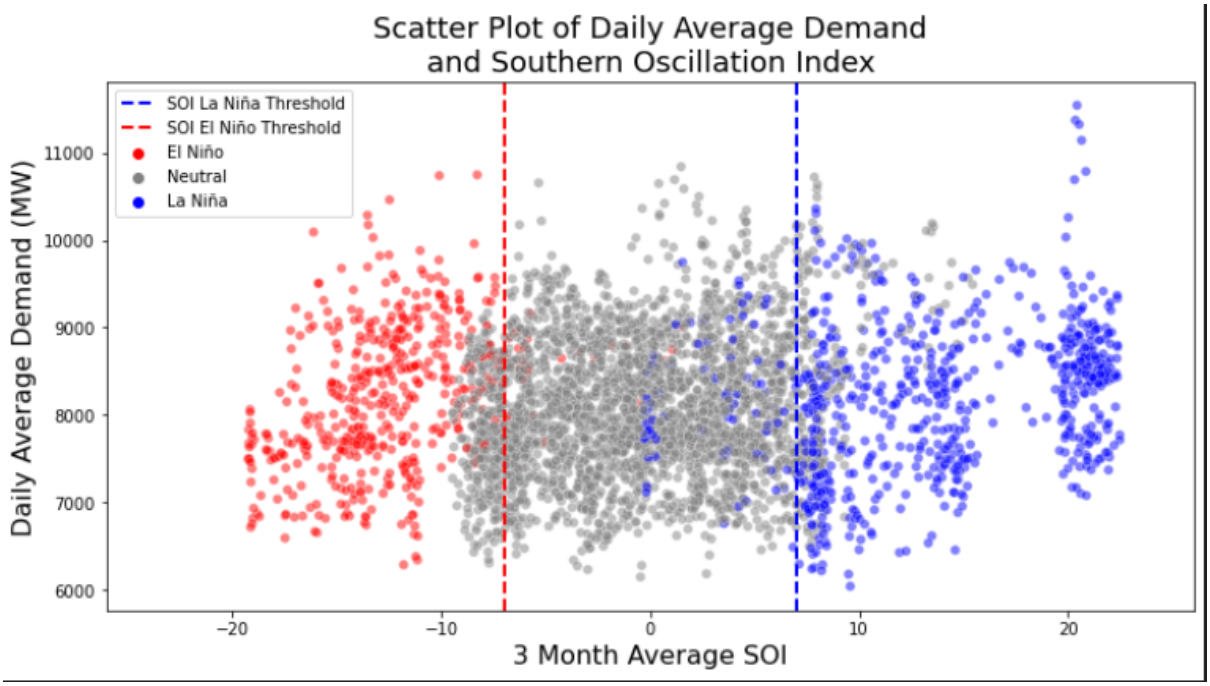


Figure 8. Scatterplot showing how SOI can be used to predict ENSO phase.

Figure 8 helps to identify how different phases of the ENSO cycle impact electricity demand. The El Niño phase appears to have greater variability in demand, while La Niña conditions lead to more stable demand typically between 7,000 MW and 9,000 MW. The neutral phase exhibits a moderate spread of demand. Understanding this relationship can aid in forecasting electricity usage based on ENSO conditions. Do note that there are some potential outliers in the La Nina phase, this suggests that during La Nina, although demand is generally stable, there are occasional spikes or drops in demand. While the correlation coefficient between actual demand and SOI is 0.044 which indicates no correlation, when processed to be the indicator of three-monthly SOI, the correlation rises to 0.129. This would indicate three-monthly SOI is one of the better linear predictors however, during the modelling process it was shown that the raw (not three-monthly) data produced better predictions for models that better handled nonlinearities.

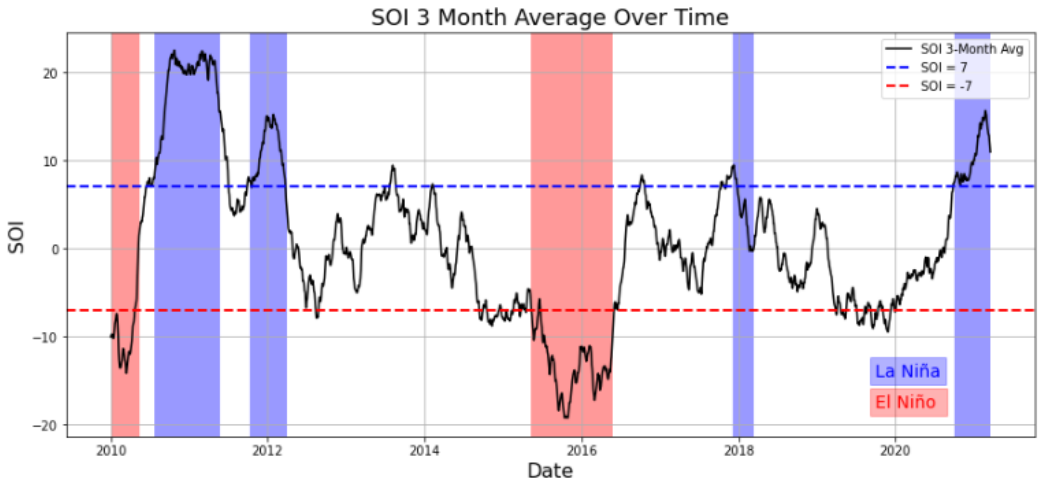


Figure 9. ENSO phase overlaid with SOI within the timeframe of the study.

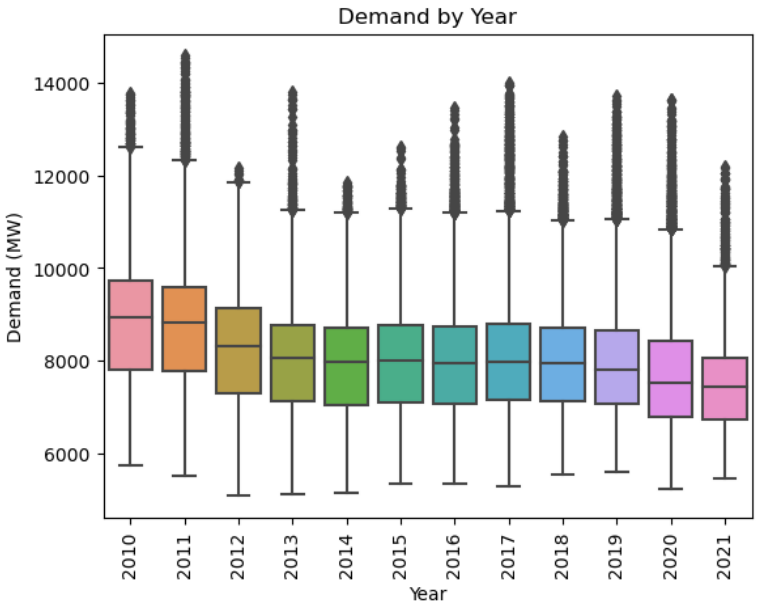


Figure 10. Boxplots of demand by year.

Comparing Figure 9 showing SOI over time and Figure 10 showing yearly demand, there seems to be slightly fewer fluctuations in demand in general during El Nino period (2015, 2016) compared to La Nina and Neutral periods. La Nina generally exhibits higher and more variable demand, however, 2012 does have the lowest fluctuations in demand in all the years we observed, therefore further investigation into additional factors would be necessary to draw firm conclusions.

4.7 Correlation between demand and features for modelling

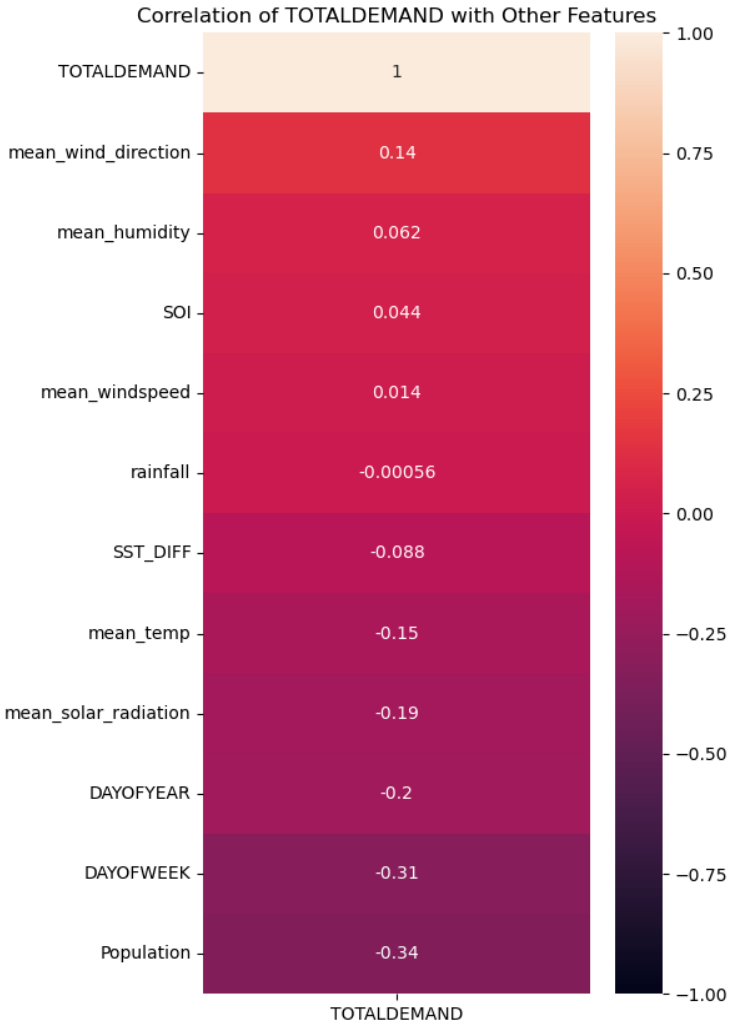


Figure 11. Heatmap of the correlation coefficients between each covariate in the dataset and demand.

- Population (-0.34): Surprisingly, as population increases, electricity demand decreases, possibly due to improvements in energy efficiency and renewable energy adoption.
- Day of the Week (-0.31): Demand tends to be lower on weekends, reflecting reduced industrial and commercial activity.
- Day of the Year (-0.2): Electricity demand shows a weak seasonal decline over the course of the year.
- Weather-related factors (e.g., temperature, humidity, solar radiation): These have weak correlations with electricity demand, indicating that weather influences are present but not dominant drivers of demand.
- SOI (0.044): ENSO cycles (SOI) show minimal direct impact on demand.

Overall, these correlations suggest that factors like temporal patterns (weekday vs. weekend) and population changes might have more influence on electricity demand than weather conditions alone, warranting further analysis. This is enough to infer that ENSO indicators have little effect on forecast energy use predictions under the assumption of linearity. However, additional observations within this exploratory data analysis indicate that there are non-linear relationships within the data which would be better modelled using an algorithm that handles non-linearity well. Note that this heat map shows only the features modelled after the appropriate ENSO metrics were selected as the raw data from which ENSO classifications are derived.

5 Analysis and Results

To benchmark all modelling results, the Root Mean Square Error (RMSE) between the forecast demand (AEMO prediction) and total demand (actual demand) was calculated and found to be 224.76 MW which is 2.7% of the mean value of the actual demand value.

5.1 Linear Regression Model

It has already been presented that there is little expectation of accurate prediction using a linear model. It is necessary to present predictions made using a linear model due to its simplicity and very low computational overhead. The data was fitted to a linear model with the variables $y = TOTALDEMAND$ and;

$$x_i = \{mean\ wind\ direction, mean\ humidity, SOI, mean\ windspeed, rainfall, SST\ difference, \\ mean\ temperature, mean\ solar\ radiation, day\ of\ the\ year, month\ of\ the\ year, population\}$$

Under the form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \beta_{10} x_{10} + \beta_{11} x_{11} + \epsilon$$

Where it is assumed $\epsilon \sim N(0,1)$.

Predictions within the time frame made using the linear model are graphed below. Train and test RMSE were found to be 385.66MW and 508.14MW respectively, which are both significantly larger than the RMSE between AEMO predictions and the actual demand. This would indicate that the model presented here fails to meet the goal of the project of improving accuracy of predictions.

These covariates x_i were used for all models to predict y in the main models, based on decisions made from initial heatmaps and feature importance graphs (Figure 20).

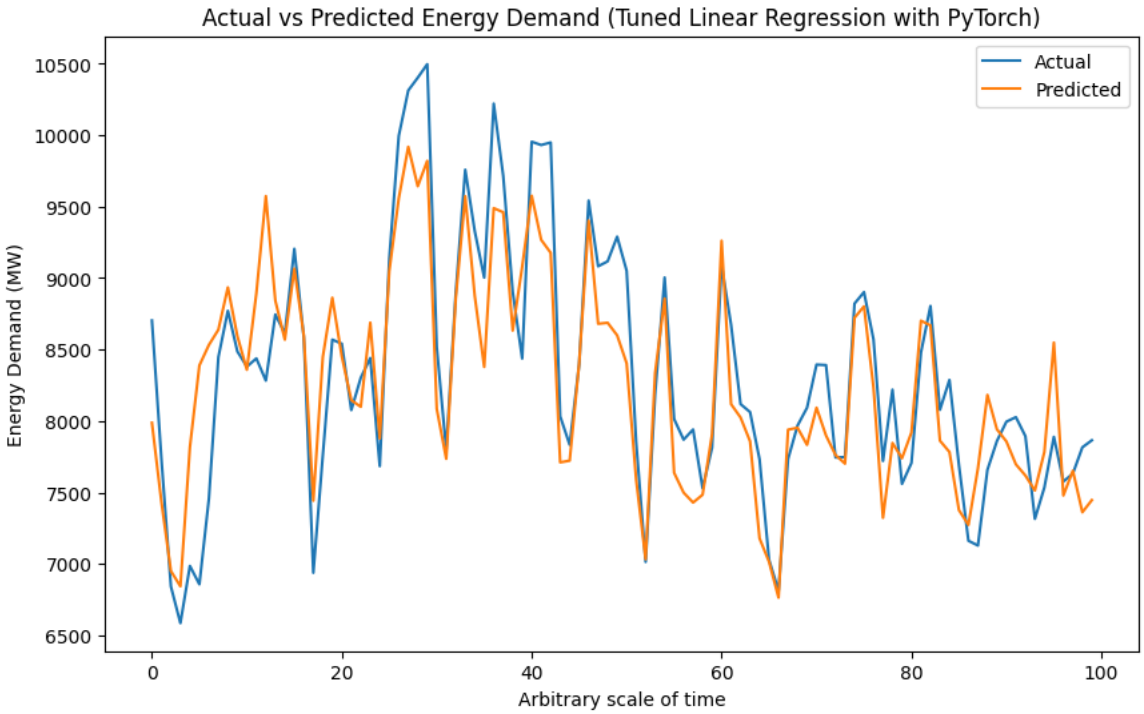


Figure 12. Comparison between predictions made using linear model to the actual data.

Note that in the linear prediction depicted in Figure 12, the model under-forecasts the demand for large spikes in demand. Rather than a general dampening of the model, it appears the model predicts drops in demand well. Obviously, for an energy producer the inability to provide enough power when it is required by the grid presents an undercapitalisation in both finance and public good.

The linear prediction still presents an advantage in computational overhead so examination of the predictions further and setting them as a benchmark for other models is necessary.

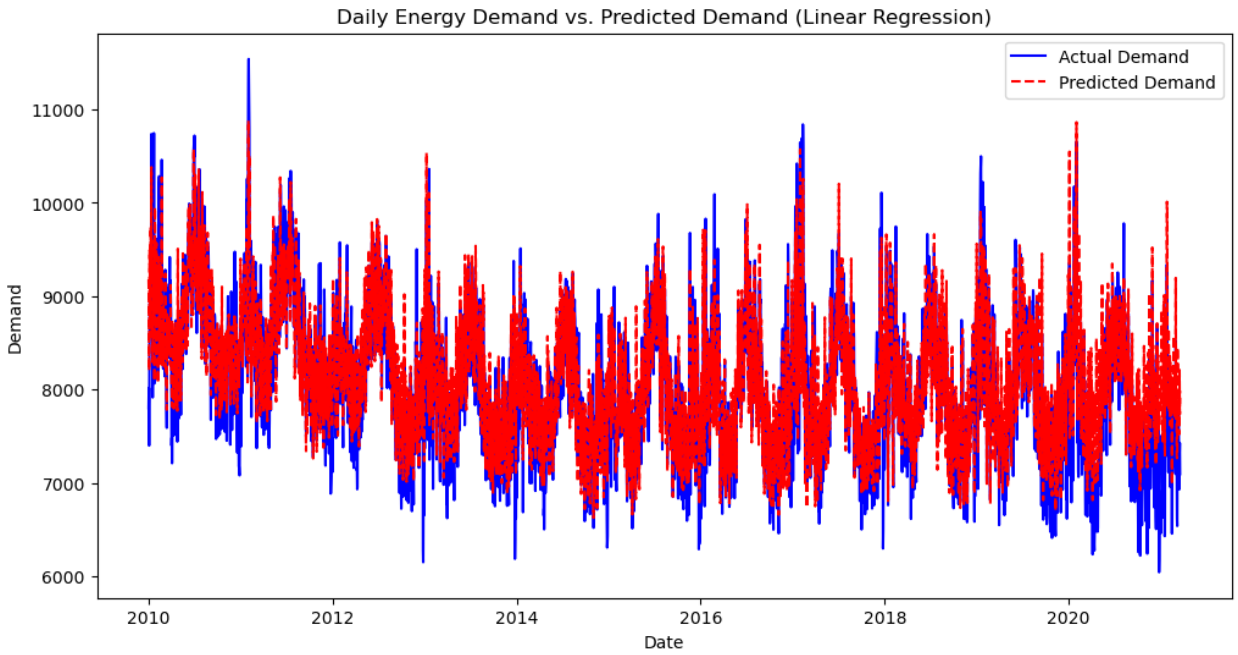


Figure 13. Validation graph of the linear prediction

Note that in the above Figure 13, there is a more apparent general dampening prediction than in Figure 12. It would therefore be irresponsible to endorse the linear regression as a good predictor even given its simplicity.

This led to consideration of the other three models explained in the introduction and literature review. Presented below are Random Forest model, XGBoost model and MLP model. Each model will have a prediction graph within the range and then prediction graphs for daily, weekly, monthly and yearly granularity until 2070 when the published population forecast ends. From those models, the best model will be selected for further investigation.

5.2 Random Forest

Random Forest produced significantly better results than the linear model. Using PyTorch and a parameter matrix, the model trained on each unique combination of parameters resulted in a Random Forest with parameters as outlined in Table 1:

Table 1. Hyperparameters of optimal Random Forest model

Number of estimators	200
Minimum samples split	2
Minimum samples in a leaf	1
Maximum features to split	1
Maximum depth	None

This yielded a model that had Train RMSE of 88.29MW and Test RMSE of 238.61MW which, while not as accurate as the AEMO predictions, is only marginally less accurate than the professional model they create.

Some advantages over the linear model observable from Figure 14 below are added sensitivity to seasonal and shorter-range fluctuations. This Random Forest model still suffers from the dampening of predictions at the peaks and troughs of demand meaning forecasts made using this model will oversupply periods of decreased demand and undersupply periods of increased demand.

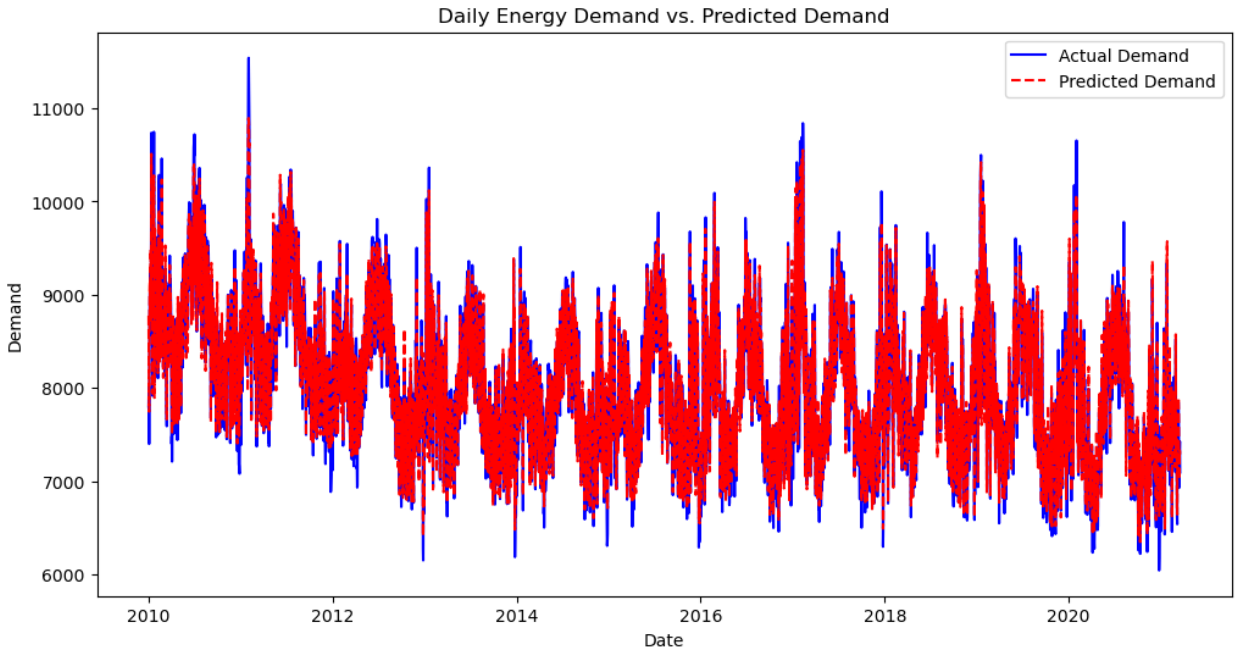


Figure 14. Predictions within the time range for the Random Forest model.

Since the Random Forest model requires lower computational overhead than an XGBoost model, if a multiplicative factor was added to these predictions to account for under- and over-predicting peaks and troughs then the Random Forest model may be the best model.

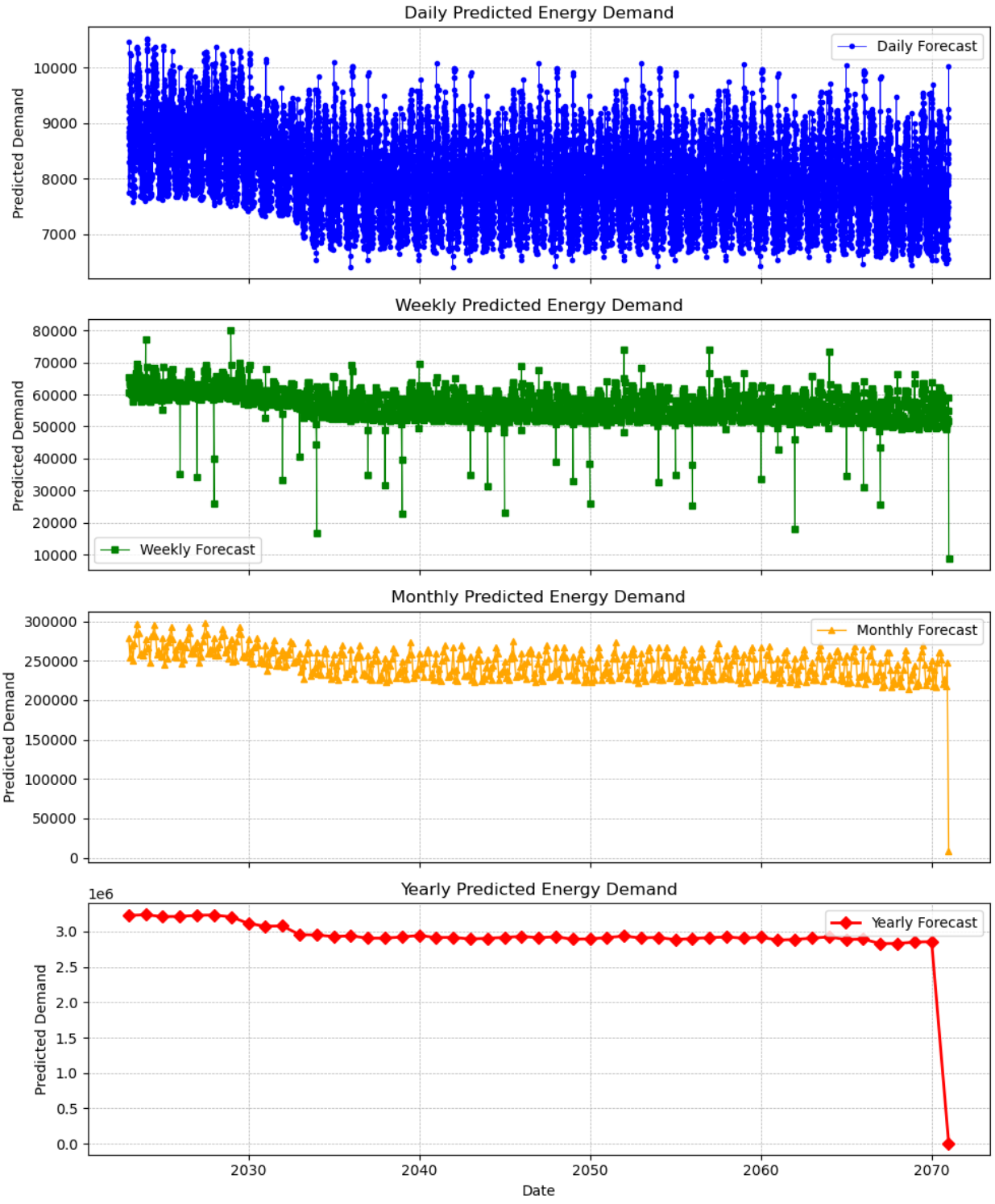


Figure 15. Daily, weekly, monthly and yearly forecasts from Random Forest model.

In Figure 15 there is a clear drop in overall predictions in the next decade for all subplots. The extreme fluctuations seen on the weekly scale would make this model (and in fact, all models presented here) inappropriate for use. The smoothness of the forecasting for monthly and yearly forecasts as well as the stability of daily forecasts means that the models

may be appropriate here. Random Forest still comes with the limitation of difficulty to account for extrapolation so the great fluctuations in the predictions may prove a problem.

5.3 XGBoost

Again, changing models to XGBoost produced better results than Random Forest. This was expected as preliminary models were used to determine the best subset of ENSO indicators. Using PyTorch and a parameter matrix, the model trained on each unique combination of parameters resulted in an XGBoost models with parameters:

Table 2. Hyperparameters of optimal XGBoost model

Subsample	0.7
Number of estimators	500
Minimum child weight	1
Maximum depth	3
Learning rate	0.1
Gamma	0.3
Colsample bytree	0.7

This yielded a model that had Train RMSE of 119.57MW and Test RMSE of 194.88MW, which outperforms AEMO predictions by more than 30MW (RMSE).

Figure 16 below shows that there is no apparent general dampening of predictions, meaning the XGBoost model accounts for the peaks and troughs of demand. This alone makes it a prime candidate for deployment, alongside its increased accuracy over the other tested models and AEMO predictions.

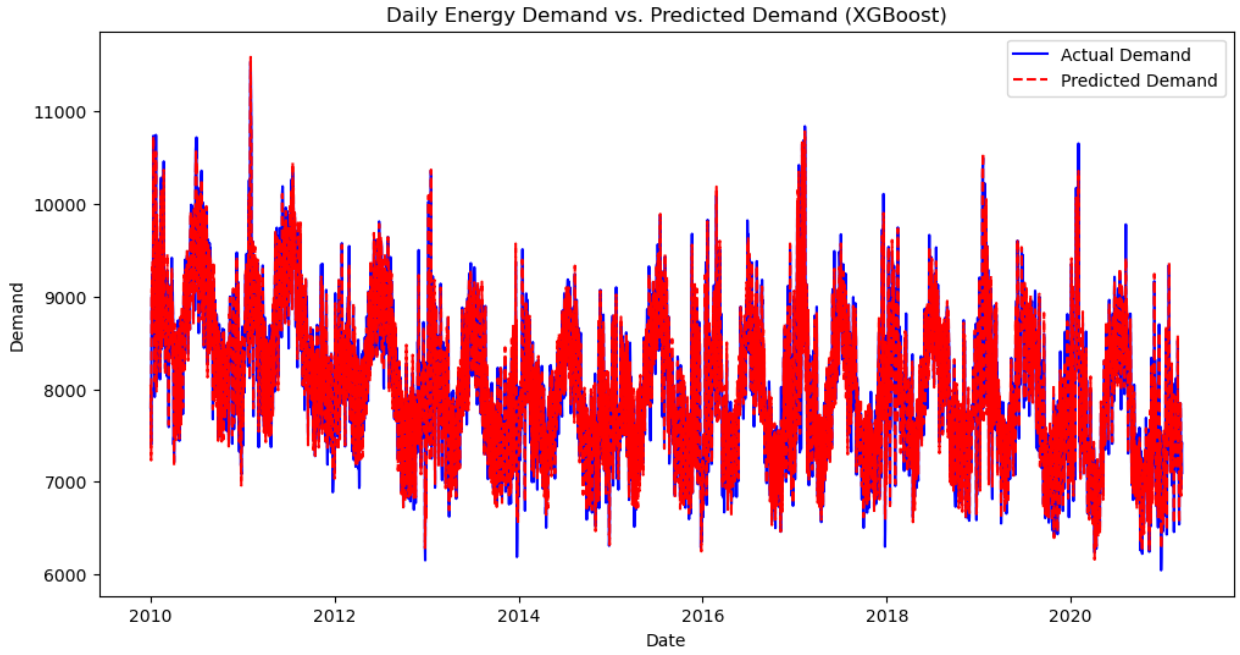


Figure 16. Predictions within the time range for the XGBoost model.

The XGBoost model had minimal impact from compute performance from the additional complexity. The XGBoost model accounts for the peaks and troughs in demand meaning that there is little chance of over- or under-supply of electricity using this model.

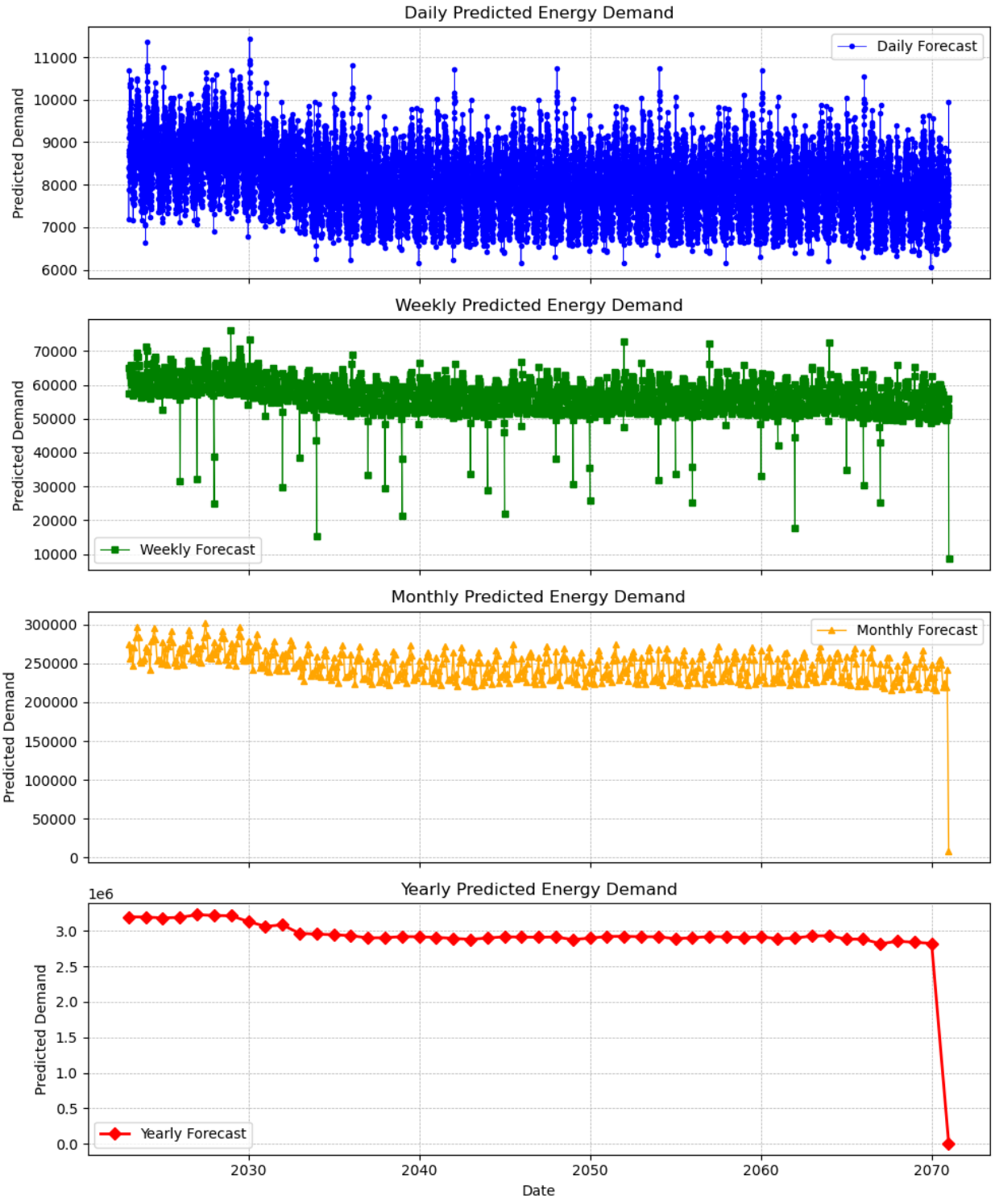


Figure 17. Daily, weekly, monthly and yearly forecasts from the XGBoost model.

The weekly predictions in Figure 17 made using the XGBoost model have similar fluctuations to the Random Forest model which similarly makes this XGBoost model inappropriate for weekly forecasts. The daily, monthly and yearly forecasts exhibit the same strengths as

those observed in the Random Forest model, however don't seem to have as dramatic a drop in 2033 that is observed for the Random Forest model. Overall, the models both have similar predictions which makes the XGBoost model the best candidate so far due to its increased accuracy.

5.4 Multilayer Perceptron Neural Network

After initial modelling and reading the literature, the MLP model was not expected to increase accuracy. There were two main reasons why the MLP was still investigated, firstly as a representative of neural networks and secondly because MLP often forms the basis of hybrid models.
Using PyTorch and a parameter matrix, the model trained on each unique combination of parameters resulted in an MLP models with parameters as shown in Table 3:

Table 3. Hyperparameters of optimal MLP model

Solver	adam
Maximum iterations	2000
Hidden layer sizes	100, 50
Learning rate initialisation	0.001
Activation	0.1
Gamma	ReLU
Random state	42

This yielded a model that had Train RMSE of 272.20MW and Test RMSE of 504.34MW which underperforms Random Forest and XGBoost and only slightly outperforms the Linear Regression model. This allows us to infer that an MLP alone is not a good predictor of energy demand.

Based on Figure 18, it appears that the MLP model improves on the Linear Regression predictions by not under-forecasting spikes in demand but would still result in over-supply of power in demand troughs.

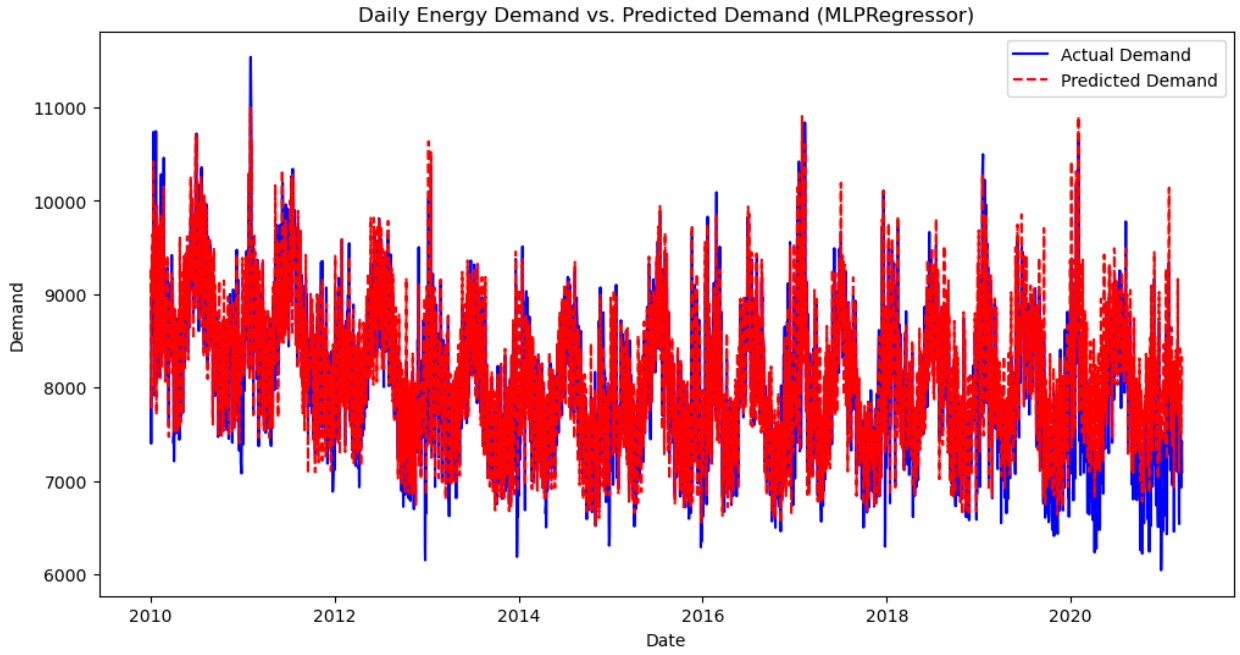


Figure 18. Predictions within the time range for the MLP model.

Since the MLP model forms the basis for many hybrid models, this is a useful result. The advantage that neural network models have over decision tree and gradient boosting models like Random Forest and XGBoost is that they work very well on extrapolation. A neural network approach would therefore be useful if the effects of a changing climate, private renewables (as opposed to grid-based) or population produce interactions that aren't captured in contemporary data.

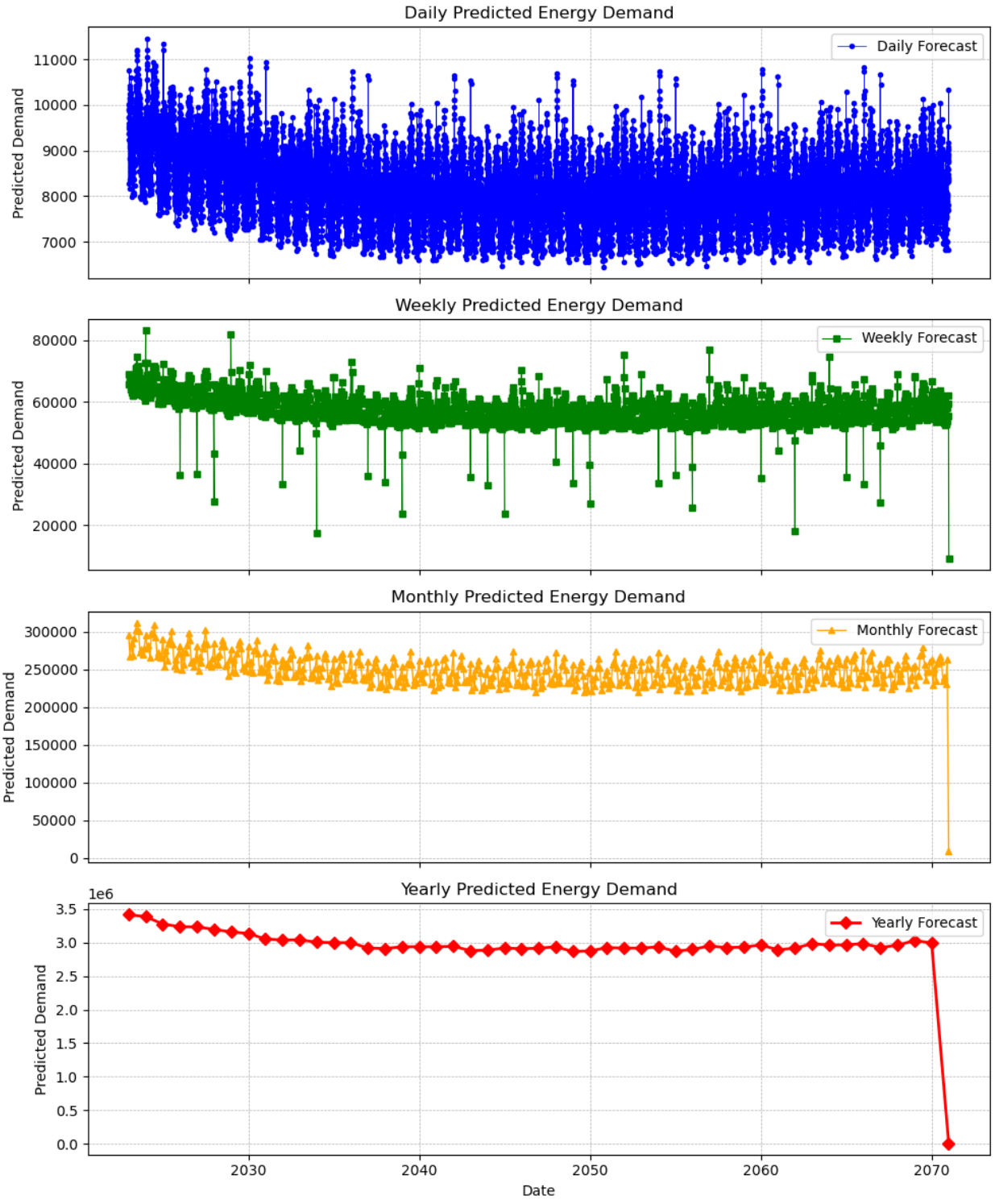


Figure 19. Daily, weekly, monthly and yearly forecasts from the XGBoost model.

Figure 19 shows the predictions made using the MLP model. As opposed to all three other models, there is no drop in the predictions around 2033. As there doesn't seem to be any basis for the model to expect a drastic difference between 2032 and 2033, the MLP model

may be the best candidate to extrapolate results from. It could therefore be recommended to look into hybrid models for a future investigation into the same topic.

5.5 Further investigation

During initial investigation and based on literature, temperature was consistently observed to have the greatest impact on energy demand forecasts. The BOM_CLASS feature had higher importance than either of its contributing SST and SOI features, however, had lower importance than the aggregate of the others. As both Random Forest and XGBoost can split the data based on feature interactions that may not be expected, it was decided that the raw data metrics of SOI and SST_DIFF would be used for all subsequent models.

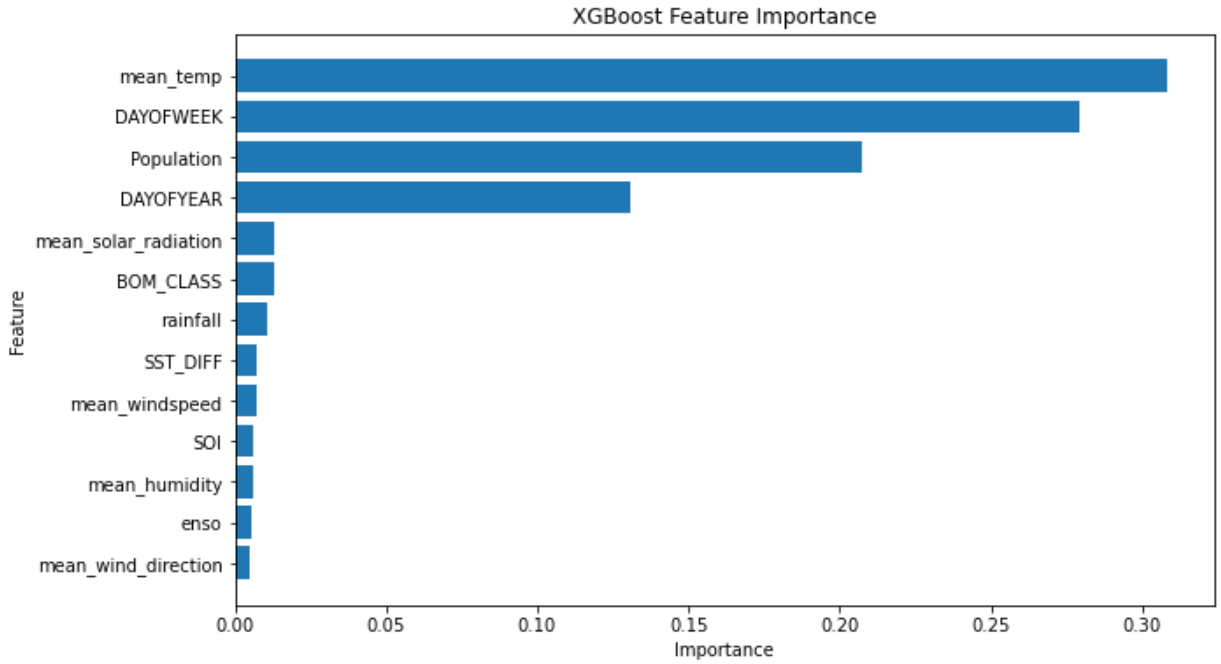


Figure 20. Feature importances of the XGBoost model for all parameters

Based on the confirmation in Figure 20 that temperature was the most important factor, the question followed as to whether a good substitute for the weather effect of temperature could be adequately substituted for the climate indicators of SST and SOI. Temperature was then dropped from the DataFrame and an XGBoost model was trained by testing the parameter space the original model had access to.

The best parameters of the model are outlined in Table 4:

Table 4. Best parameters of the XGBoost model without temperature

Subsample	0.9
Number of estimators	500
Minimum child weight	3
Maximum depth	3
Learning rate	0.2
Gamma	0.5
Colsample bytree	0.7

This resulted in a model with a training RMSE of 162.79MW and test RMSE of 341.80MW, which are significantly worse than the same model with temperature included. This model does not accurately forecast the energy demand within the range and exhibits the general dampening observed in the Random Forest and Linear Regression models. This can be observed from Figure 21.

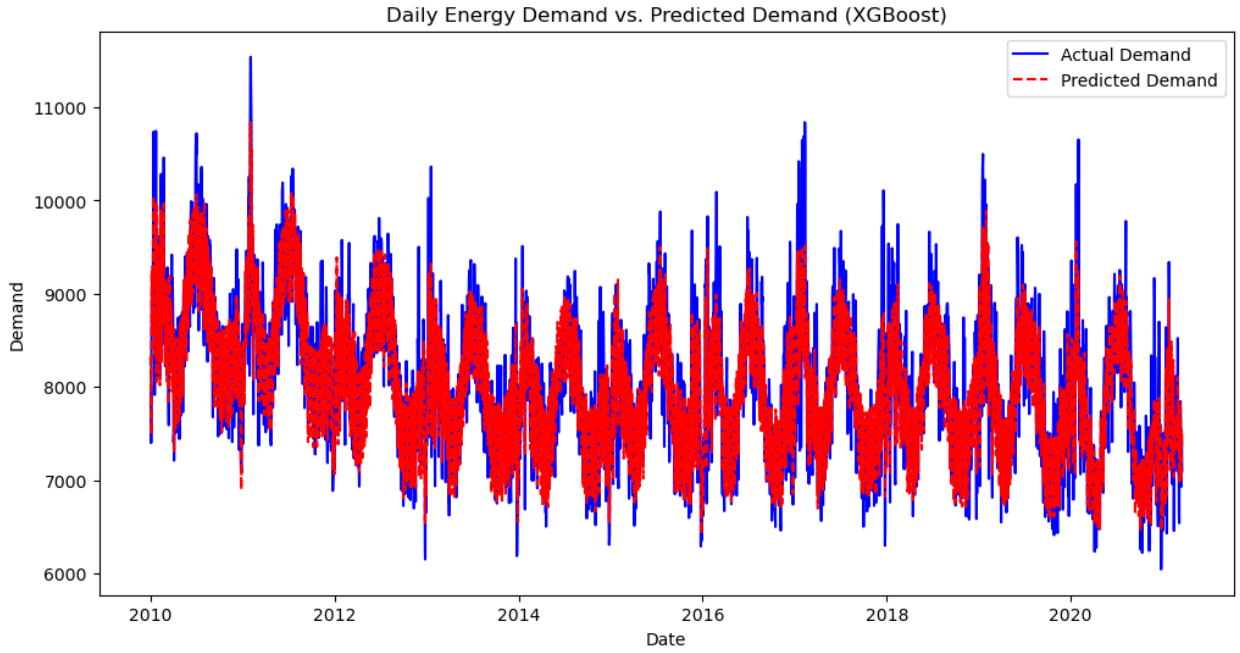


Figure 21. Predictions within the time range for the XGBoost model without temperature.

Like all other models exhibiting the dampening and without the advantages in computational overhead of Random Forest and Linear Regression, the XGBoost model without temperature is not a viable candidate to make accurate predictions on energy demand.

Interestingly from feature importance, the day of the year becomes more important than population when temperature cannot be accounted for as shown in Figure 22. This leads to the assumption that since each day of the year is inextricably linked to its historical temperature, it would be a fair correlary to estimate temperature from the day of the year. It is also of note that the aggregate feature importance of SST_DIFF and SOI would be 0.052299, placing their performance between mean_solar_radiation and Population but still insignificant in comparison to the three main factors.

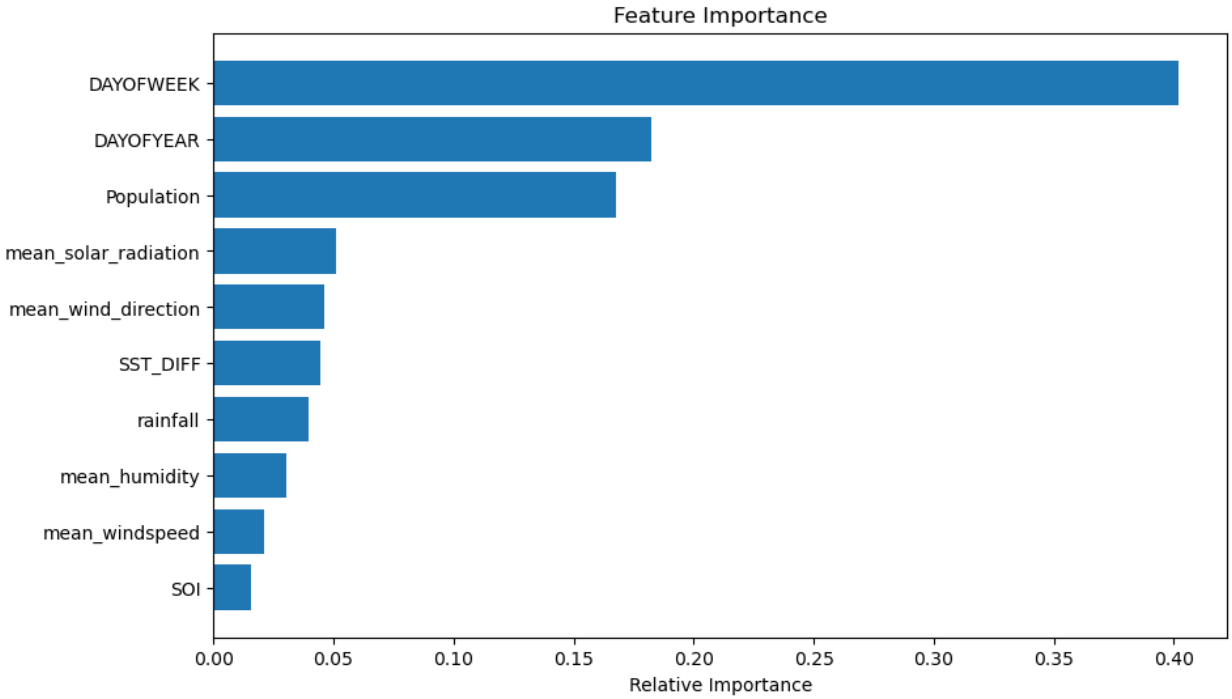


Figure 22. Feature importance of best XGBoost model without temperature or superfluous ENSO candidates

Whilst the further investigation of XGBoost without temperature yielded interesting results, it does not produce a useful alternative to using temperature as the main forecasting feature. It would therefore be sensible to conclude that XGBoost with the full suite of variables is the best model to forecast energy demand on the daily, monthly and yearly scale. None of the models performed very well as a weekly predictor and therefore a weekly forecasting paradigm cannot be produced from these results.

6 Discussion

All four models selected within this study showed ability to forecast accurately within the realms of viability for daily, monthly and yearly predictions. Only Random Forest (238.61MW) and XGBoost (194.88MW) were able to approach a similar accuracy (RMSE) of the AEMO predictions of 224.76MW. Of these, Random Forest was unable to account for extreme fluctuations in demand, meaning that spikes in demand would be under-forecast by the Random Forest model and troughs in demand would be over-forecast by the Random Forest model. In the case of under-supply, this would result in undercapitalisation of the market opportunity by not producing enough electricity and in the converse case, the undercapitalisation would be the result of burning unnecessary fossil fuels that would saturate the grid. Grid saturation would have a further effect of potentially powering battery supply operators, giving competitors a market edge.

XGBoost when tuned appropriately, was able to outperform the AEMO model on the given data range. This was unexpected for a model trained on historical data however, this study has the advantage of hindsight in that the models could be tuned to exceed AEMO accuracy. This is in contrast to AEMO which must produce forecasts in real time based on the metrics and models they use and are unable to retroactively update their forecasts to better reflect the actual demand. The XGBoost model performing better on this dataset does not imply poor modelling on the behalf of AEMO but only that there are advantages to the conditions under which a report must be constructed.

Since the scale of forecast most important to a legacy energy producer is the monthly demand, computational overheads are less of a concern so long as forecasts can be made within a reasonable timeframe. Since the XGBoost model took 22.8 seconds to train and the Random Forest model took 51.5 seconds to train on a consumer MacBook, it is easy to conclude that the computational overhead of these models is no downside when forecasting on this time scale.

The inclusion of El Niño and La Niña metrics in the feature set for these models is of nebulous use. When temperature is removed, they seem to be of greater significance but still do not play a vital role in forecasting.

While the recommendation of this report will be to use the XGBoost model, Figures 15, 17 and 19 all demonstrate that a monthly forecast, which would be used by legacy energy producers to stock the fossil fuel they require to operate, is relatively stable and consistent for any of the models investigated. All four models are in good agreement overall and it would be appropriate to use any if one of the disadvantages of XGBoost was unsatisfactory. As examples, Linear Regression may be chosen for its simplicity or MLP may be chosen for its malleability and ability to handle extrapolation well.

7 Conclusion and Further Issues

Machine learning models were used to determine if the addition of ENSO metrics as features would improve performance of energy demand forecasts when compared to those produced by AEMO. The techniques used were Linear Regression, Random Forest, XGBoost and Multilayer Perceptron Neural Network.

The RMSE of the AEMO's forecasted demand being approximately 225MW was bettered by XGBoost, which was the best performing model with a test RMSE of approximately 195MW. This highlights the strength of this machine learning algorithm in the setting of forecasting energy demand. All four models provided valuable monthly forecasts however Random Forest and XGBoost performed significantly better overall than Linear Regression and Multilayer Perceptron. The limitation of using XGBoost for forecasting into the far future (which we have done here) is that it does not account well for extrapolation when features fall outside of the range of the training set. If this were to happen, which would be if temperatures were significantly higher or demand became significantly lower, the model may not adequately respond. If this were to happen or if it were expected to happen, then a hybrid approach based on MLP would be preferred as it accounts for non-linearity and seasonality but requires improved accuracy.

The key predictors of demand crucial to the success of these models were temperature, day of the week, population, and day of the year. Meanwhile, the inclusion of ENSO data was found to have no significant effect on the performance of any models. It can be concluded then that ENSO metrics do not provide any valuable improvement to models. All performance beyond the AEMO forecasts is therefore only attributed to the tuning of the XGBoost using weather metrics already captured by AEMO.

While it was assumed that New South Wales is a good substitute for the case of ENSO in any Eastern state, it may be of use to approach the same problem for the North-Eastern state of Queensland due to its greater exposure to tropical weather patterns. The models considered in this study represent an entry point into machine learning models and neural networks. Applying hybrid models, Convolutional Neural Networks, LSTM or other more advanced models may also produce more accurate and valid models.

References

Afzal, S. et al. (2023) 'Building energy consumption prediction using multilayer perceptron neural network-assisted models; comparison of different optimization algorithms', Energy, 282. Available at: https://doi.org/10.1016/j.energy.2023.128446.

AIML maintainers (no date) What are the advantages and disadvantages of Random Forest?, AIML.com.

Albuquerque, P.C., Cajueiro, D.O. and Rossi, M.D.C. (2022) 'Machine learning models for forecasting power electricity consumption using a high dimensional dataset', Expert Systems with Applications, 187. Available at: https://doi.org/10.1016/j.eswa.2021.115917.

Australian Energy Market Operator (2022) Forecasting Approach-Electricity Demand Forecasting Methodology.

Australian Photovoltaic Institute (2024) State Performance, APVI.

Bedi, J. and Toshniwal, D. (2019) 'Deep learning framework to forecast electricity demand', Applied Energy, 238, pp. 1312–1326. Available at: https://doi.org/10.1016/j.apenergy.2019.01.113.

Bilgili, M. and Pinar, E. (2023) 'Gross electricity consumption forecasting using LSTM and SARIMA approaches: A case study of Türkiye', Energy, 284. Available at: https://doi.org/10.1016/j.energy.2023.128575.

Chreng, K., Lee, H.S. and Tuy, S. (2022) 'A Hybrid Model for Electricity Demand Forecast Using Improved Ensemble Empirical Mode Decomposition and Recurrent Neural Networks with ERA5 Climate Variables', Energies, 15(19). Available at: https://doi.org/10.3390/en15197434.

Eskeland, G.S. and Mideksa, T.K. (2010) 'Electricity demand in a changing climate', Mitigation and Adaptation Strategies for Global Change, 15(8), pp. 877–897. Available at: https://doi.org/10.1007/s11027-010-9246-x.

Fan, H., Macgill, I.F. and Sproul, A.B. (2015) Statistical analysis of driving factors of residential energy demand in the greater Sydney region, Australia. Available at: https://doi.org/http://dx.doi.org/10.1016/j.enbuild.2015.07.030.

Fan, H., MacGill, I.F. and Sproul, A.B. (2017) 'Statistical analysis of drivers of residential peak electricity demand', Energy and Buildings, 141, pp. 205–217. Available at: https://doi.org/10.1016/j.enbuild.2017.02.030.

Fan, S. and Hyndman, R.J. (2010) 'Short-term load forecasting based on a semi-parametric additive model', IEEE Transactions on Power Systems. Melbourne. Available at: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://robjhyndman.com/papers/2010STLF-FinalR1.pdf (Accessed: 17 September 2024).

Fattah, J. et al. (2018) 'Forecasting of demand using ARIMA model', International Journal of Engineering Business Management, 10. Available at: https://doi.org/10.1177/1847979018808673.

International Energy Agency (2023) Keeping cool in a hotter world is using more energy, making efficiency more important than ever, IEA.

Koprinska, I., Wu, D. and Wang, Z. (2018) 'Convolutional Neural Networks for Energy Time Series Forecasting', in 2018 International Joint Conference on Neural Networks (IJCNN). IEEE, pp. 1–8. Available at: https://doi.org/10.1109/IJCNN.2018.8489399.

Kumar Dubey, A. et al. (2021) 'Study and analysis of SARIMA and LSTM in forecasting time series data', Sustainable Energy Technologies and Assessments, 47. Available at: https://doi.org/10.1016/j.seta.2021.101474.

Leung, A. (2022) 'A Multivariate Model for Electricity Demand using Facebook Prophet', TowardsDataScience [Preprint].

Liu, Y. et al. (2023) 'Enhanced multi-year predictability after El Niño and La Niña events', Nature Communications, 14(1). Available at: https://doi.org/10.1038/s41467-023-42113-9.

Mahjoub, S. et al. (2022) 'Predicting Energy Consumption Using LSTM, Multi-Layer GRU and Drop-GRU Neural Networks', Sensors, 22(11). Available at: https://doi.org/10.3390/s22114062.

Mavuduru, A. (2020) Why XGBoost Can't Solve All Your Problems, TowardsDataScience.
Mfetoum, I.M. et al. (2024) 'A multilayer perceptron neural network approach for optimizing solar irradiance forecasting in Central Africa with meteorological insights', Scientific Reports, 14(1). Available at: https://doi.org/10.1038/s41598-024-54181-y.

Nooruldeen, O. et al. (2023) 'Strategies for predictive power: Machine learning models in city-scale load forecasting', e-Prime - Advances in Electrical Engineering, Electronics and Energy, 6. Available at: https://doi.org/10.1016/j.prime.2023.100392.

Porteiro, R., Hernández-Callejo, L. and Nesmachnow, S. (2022) 'Electricity demand forecasting in industrial and residential facilities using ensemble machine learning', Revista Facultad de Ingeniería, Universidad de Antioquia [Preprint]. Available at: https://doi.org/https://www.doi. org/10.17533/udea. redin.20200584.

Qureshi, M., Arbab, M.A. and Rehman, S. ur (2024) 'Deep learning-based forecasting of electricity consumption', Scientific Reports, 14(1). Available at: https://doi.org/10.1038/s41598-024-56602-4.

Scikit Learn Maintainers (2024a) Compare Stochastic learning strategies for MLPClassifier, Scikit Learn Documentation.

Scikit Learn Maintainers (2024b) MLPClassifier, Scikit Learn Documentation.

Thompson, B. (2019) A limitation of Random Forest Regression, Towards Data Science.

Vijendar Reddy, G. et al. (2023) 'Electricity Consumption Prediction Using Machine Learning', in E3S Web of Conferences. EDP Sciences. Available at: https://doi.org/10.1051/e3sconf/202339101048.

Vu, D.H., Muttaqi, K.M. and Agalgaonkar, A.P. (2015) 'A variance inflation factor and backward elimination based robust regression model for forecasting monthly electricity demand using climatic variables', Applied Energy, 140, pp. 385–394. Available at: https://doi.org/10.1016/j.apenergy.2014.12.011.

Zhou, H.S. et al. (2021) 'A case study on the behaviour of residential battery energy storage systems during network demand peaks', Renewable Energy, 180, pp. 712–724. Available at: https://doi.org/10.1016/j.renene.2021.08.107.

Appendix

Data Sources

Weather Data – https://www.airquality.nsw.gov.au/air-quality-data-services/data-download-facility

ENSO Criteria - http://www.bom.gov.au/climate/updates/articles/a008-el-nino-and-australia.shtml

ENSO Data - http://www.bom.gov.au/climate/enso/indices.shtml

Alternate ENSO Data (not used) - https://www.longpaddock.qld.gov.au/soi/soi-data-files/

Temperature and Demand Data - https://github.com/UNSW-ZZSC9020/project

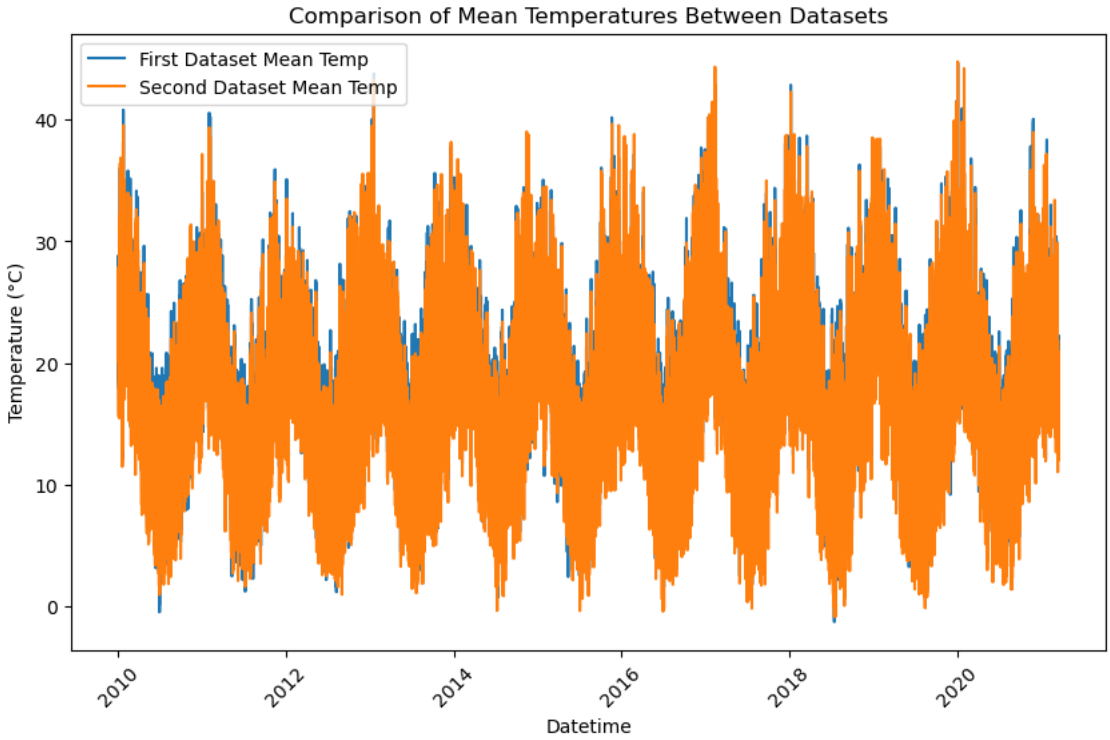
Codes

```
import pandas as pd

#enso
enso = pd.read_csv('data/enso/daily_enso.csv', header=0)
enso['Date'] = pd.to_datetime(enso['DATETIME'])
enso.drop(columns=['DATETIME','enso'], inplace=True)


#humidity
hum = pd.read_csv('data/NSW/aggregated_humidity_data.csv', header=0)
hum['Date'] = pd.to_datetime(hum['Date'], dayfirst=True)
hum.drop(columns= ['Time','median_humidity'], inplace=True)
hum = hum.groupby('Date').mean()
hum = hum.sort_values(by='Date', ascending=True)

#radiation
rad = pd.read_csv('data/NSW/aggregated_solar_radiation_data.csv', header=0)
rad['Date'] = pd.to_datetime(rad['Date'], dayfirst=True)
rad.drop(columns= ['Time','median_solar_radiation'], inplace=True)
rad = rad.groupby('Date').mean()
rad = rad.sort_values(by='Date', ascending=True)

#temp
temp = pd.read_csv('data/NSW/aggregated_temperature_data.csv', header=0)
temp['Date'] = pd.to_datetime(temp['Date'], dayfirst=True)
temp.drop(columns= ['Time','median_temp'], inplace=True)
temp = temp.groupby('Date').mean()
temp = temp.sort_values(by='Date', ascending=True)

#windtheta
```

```
windtheta = pd.read_csv('data/NSW/aggregated_wind_direction_data.csv', header=0)
windtheta['Date'] = pd.to_datetime(windtheta['Date'], dayfirst=True)
windtheta.drop(columns= ['Time','median_wind_direction'], inplace=True)
windtheta = windtheta.groupby('Date').mean()
windtheta = windtheta.sort_values(by='Date', ascending=True)

#windspeed
windspeed = pd.read_csv('data/NSW/aggregated_windspeed_data.csv', header=0)
windspeed['Date'] = pd.to_datetime(windspeed['Date'], dayfirst=True)
windspeed.drop(columns= ['Time','median_windspeed'], inplace=True)
windspeed = windspeed.groupby('Date').mean()
windspeed = windspeed.sort_values(by='Date', ascending=True)

#rainfall
rain = pd.read_csv('data/NSW/median_rainfall_2010_2021.csv', header=0)
rain['Date'] = pd.to_datetime(rain['Date'])
rain.rename(columns={'Rainfall amount (millimetres)':'rainfall'}, inplace=True)
rain = rain.sort_values(by='Date', ascending=True)

#pop
pop = pd.read_csv('data/Population/daily_pop_nsw.csv', header=0)
pop['Date'] = pd.to_datetime(pop['Date'])
pop = pop.sort_values(by='Date', ascending=True)

#TOTALDEMAND
demand = pd.read_csv('data/NSW/totaldemand_nsw.csv', header=0)
demand['Date'] = pd.to_datetime(demand['DATETIME'], dayfirst=True)
demand['Date'] = demand['Date'].dt.date
demand['Date'] = pd.to_datetime(demand['Date'])
demand.drop(columns=['DATETIME'], inplace=True)
demand = demand.groupby('Date').mean()
demand = demand.sort_values(by='Date', ascending=True)


#join the dataframes
join = pd.merge(hum, enso, on='Date', how='left')
join = pd.merge(join, rad, on='Date', how='left')
join = pd.merge(join, temp, on='Date', how='left')
join = pd.merge(join, windtheta, on='Date', how='left')
join = pd.merge(join, windspeed, on='Date', how='left')
join = pd.merge(join, rain, on='Date', how='left')
join = pd.merge(join, pop, on='Date', how='left')
join = pd.merge(join, demand, on='Date', how='left')

join['DAYOFWEEK'] = join['Date'].dt.weekday
join['DAYOFYEAR'] = join['Date'].dt.dayofyear

#drop missing values
```

```
join = join.dropna()

#save as data_for_ml.csv
join.to_csv('data/NSW/data_for_ml.csv', index=False)
```

Appendix 1. join_data_for_ml.py



Appendix 2. Temperature data overlays for the first (Bankstown Airport) and second (Sydney South-West) datasets to test assumptions about temperature.

```
param_distributions = {
    'n_estimators': [100, 200, 300, 500],
    'max_depth': [10, 20, 30, None],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4],
    'max_features': [1.0, 'sqrt', 'log2']  # Remove 'auto', use 1.0 as default or
other options
}
```

Appendix 3. Parameter space for Random Forest Model

```
param_distributions = {
```

```
    'n_estimators': [100, 200, 300, 500],
    'max_depth': [3, 6, 10, 15],
    'learning_rate': [0.01, 0.1, 0.2],
    'subsample': [0.7, 0.8, 0.9, 1.0],
    'colsample_bytree': [0.5, 0.7, 0.9, 1.0],
    'min_child_weight': [1, 3, 5],
    'gamma': [0, 0.1, 0.3, 0.5]
}
```

Appendix 4. Parameter space for XGBoost model

```
model = MLPRegressor(
    solver='adam',              # Changed to 'adam' for better convergence
    max_iter=2000,              # Increased max iterations
    hidden_layer_sizes=(100, 50),  # Adjusted hidden layer sizes
    learning_rate_init=0.001,   # Lower initial learning rate
    activation='relu',          # Changed activation function to ReLU
    random_state=42
)
```

Appendix 5. Parameter space for MLP model