32252 Data Science Group Project 2023-2024

How do weather patterns correlate with per capita electricity consumption in the UK, considering the distinctions that can be observed in the impact on electricity derived from various sources using datasets spanning from 1995 to 2021?

Mohamad Abdallah, Ali Abouyahia, Sam Robbins, Amanjot Singh

April 25, 2024

Abstract

Weather and energy exhibit strong connections; attempts have been made to identify these relationships. More precisely how is electricity consumption per capita from different sources greatly affected by weather factors such as seasonal changes and extreme weather conditions. In the past research into factors that affect electricity consumption mostly missed taking weather into account; leaving a gap in the literature. This project studies which electricity sources are most affected by changes in weather and the potential to model this effectively. The weather sources being investigated are temperature (C), rainfall (mm), wind (Knots) and sunshine (Hours). Multiple datasets from a variety of sources were cleaned and merged into a single dataset in order to work more efficiently. This was followed by a thorough Exploratory Data Analysis (EDA) in both SQL and Python. The data was subsequently examined, providing valuable insights into potential patterns and correlations. Machine learning algorithms including multiple linear regression (support vector regression), random forest, neural networks, and XGBoost, were trained and tested on this data. Evaluation of the models returned metrics showing model effectiveness. There was a large variation in model effectiveness with the random forest regression model producing the most consistently accurate performance. Hydro electricity was the most effective target to model showing an extremely high dependence on weather conditions. Our research shows that weather constituents can be used to predict electricity consumption of different sources, but due to various limitations not all characteristics of electricity consumption can be modelled accurately. To extend the study in future, the addition of other features such as import and export or economic influence in combination with fine tuning the applied models, could lead to more consistent and effective model performance.

Keywords : *Energy* • *Weather* • *Electricity Consumption* • *Machine Learning*

1 Introduction

Weather conditions leads to variations in electricity consumption due to the affects on heating or cooling requirements. In the UK, heating demands have the greatest impact since it's uncommon for residences to be equipped with air conditioning for cooling during the summer. Consequently, extensive studies have been conducted to explore fluctuations in electricity consumption and weather conditions. These have been aided with the introduction of modelling through machine learning algorithms, which allows for developing smart energy management, understanding environmental impacts, and playing a pivotal role in formulating effective sales strategies. With the advancement of machine learning algorithm, it has become common for organisation to use various forecasting techniques for prediction such as support vector machine, extreme gradient boosting, artificial neural networks and random forest regression models[1, 2, 3, 4, 5, 6, 7, 8, 9]. Previously, weather constituents were explored against total energy consumption disregarding contributions of individual electricity sources. This gap in previous research is fulfilled by exploring multiple weather patterns of rainfall, sunshine duration, temperature, and wind speed against consumption of different electricity sources with a distinction between renewable and non-renewable, which are total fuel used, coal, oil, gas, nuclear, solar, wind, hydro, and bioenergy. Thus, this research aligns with the previous studies which are further explored in the report. This study explores the use of machine learning models using various weather features to find implications on consumption of different types of electricity sources in the UK. This aims to expand the knowledge in the energy market and the historical fluctuations in the energy consumption. These fluctuations might come from factors of a decline in manufacturing and increase of population growth [10, 11].

2 Literature Review

This section unveils the unique aspects of our research, delving into related work and the factors that have been previously considered [12, 13, 2, 14, 15]. By narrowing down the focus to weather patterns such as rainfall, sunshine, temperature, and wind, our new approach scrutinises the individual impacted sources of electricity consumption per capita from these weather patterns, providing insights into which sectors should be prioritised in the UK. The energy sector typically refers to studies conducted over a short-term, medium-term, and long-term.

Regarding weather-related components, previous studies show that temperature is the most influential weather factor on electricity consumption. With increasing average temperatures over the years, we can see a decreasing load during the winter. Wind speed also affects the need for cooling or heating. Rainfall is location-dependent and affects regional domestic consumption. Sunshine is also a significant factor as it estimates how much cloud coverage there is and how it may affect some renewable sources as well as fluctuations in energy requirements. We found correlation factors between energy sources and weather, for example hydro with correlation coefficients of -0.62, -0.66, 0.53 and 0.28 for temperature, sunshine, rainfall and wind respectively. The negative correlations between temperature and electricity consumption are due to the UK's non-dependence on air conditioning for cooling, so in the summer, temperatures rise, and electricity consumption decreases since they are not using heating anymore. These correlation factors indicate a strong relationship between weather patterns and hydroelectric power generation, which is a key aspect of our research on electricity consumption.

Traditional methods of forecasting include regression models, exponential smoothing models, and time series models [16, 17, 18]. Although we can see a monthly electricity consumption trend, with the effects of climate change and other economic factors, the data exhibits an overall growth and volatility, which, with traditional modelling methods, have resulted in poor performance because of these.

At present, self-supervised learning algorithms have been proven to achieve results that not only match but often surpass those obtained by the previously mentioned methods. These methods, such as support vector machine, extreme gradient boosting (XGBoost), artificial neural networks (ANN), and random forest regressors or decision tree based approaches, are highly reliable, and have been extensively tested on similar datasets, instilling confidence in their effectiveness [1, 2, 3, 4, 5, 6, 7, 8, 9].

A short-term forecast of electricity consumption conducted by Fazil Kaytez et al. (2015) for Turkey conducts a comparative study of multiple linear regressions (MLP), artificial neural network (ANN), and least-square support vector machine (LS-SVM), found that the LS-SVM tuned model outperformed the rest with an R-squared of 99.98% on their test results [1]. Although the LS-SVM is a new formulation of regular SVM and is computationally more efficient than the standard version, its parameters are significantly more challenging to tune [19].

A study by T. Zhang et al. (2023) for the New South Wales in the UK power network combines sequential configurations with the XGBoost algorithm to achieve long-term energy and peak power research [3]. The results found a mean absolute percentage error (MAPE) of 1.93%, which is better than the compared results with ANNs. Although the study also shows that XGBoost was more reliable in predicting per-season electricity consumption, proving its robustness vs outliers, it is prone to overfitting and is memory intensive.

N. Jaisumroum et al. (2017) performed both ANN and multiple linear regression (MLR) on the data from the Electricity Generating Authority of Thailand for different energy sources from 1993 to 2015 [7]. The variables used for predictions include gross domestic product (GDP), population, gross electricity generation, and installed capacity. These are essential features in studying electricity consumption, which most previous papers mentioned use. The R-squared from ANN of 99.98% was superior to MLR with 95.55% over the total electricity consumption. Since the amount of data used for this dataset is small, the ANN is more prone to overfitting and is more data-dependent compared to MLR, which does not address the complexity of the patterns.

Pang X et al. (2022) use a random forest regressor to determine monthly electricity consumption based on the maximum mutual information coefficient value [9]. The smart method in this paper utilizes the coefficient to provide bigger weights on impacting factors and then uses the random forest for predictions of electricity consumption in Shenyang from

2005 to 2015. As a benchmark, it is compared to the widely used SVM for predictions; the proposed method showed an improvement of 7.29% over the mean average percentage error (MAPE), putting the random forest method as a top competitor in the industry. These findings have significant implications for the energy sector, providing a reliable tool for monthly electricity consumption forecasting, thereby emphasising the practical relevance of our research.

Our paper aims to show that we may obtain similar results with the top methods used. As seen from the results, random forest is the top competitor. This finding has significant implications for the energy sector, as it provides a reliable tool for monthly electricity consumption forecasting. The UK needs more research regarding a per-source comparison, which we have provided using state-of-the-art methods. This demonstrates the potential of self-supervised learning algorithms in improving our understanding of weather patterns and their impact on electricity consumption.

3 Question Development

The total electricity demand has significantly dropped since the 1970s, mainly provoked by the displacement of industries to different countries [20]. Despite efforts to reduce carbon emissions, from 1.4 million tones of oil equivalent (Mtoe) to 142.0 Mtoe, the electricity market in the UK is experiencing increased fragmentation and competition, moving away from a monopolistic structure [21]. The fragmentation severely impacts the reliability of studying the entire market, so we study the impact of various weather patterns with each recourse, which gives us a fresh perspective on where to conduct further studies. Research showed that most publications, as discussed previously, oversee many influential variables impacting the market but only study overall consumption. There needs to be an energy-specific focus in the literature for those who want to study sector-focused studies on these different recourses, especially in the UK. With intelligent grid systems showing up in different parts of the world, we now more than ever need to understand how to manage it and correctly price it to the market. This paper is the UK's first time studying per source for weather patterns. Increased temperatures and varying weather patterns significantly impact the capacity of transmission and distribution lines, which leaves us with the need to continuously further our studies in this market [22]. Since the UK has publicly accessible data for monthly electricity and weather data, it was grouped by the team, making it reliable to study the UK as a whole for each resource.

4 Methodology

4.1 Data Collection

Multiple datasets have been utilised to fill the gap in the literature, with a focus on utilising monthly data for prolonged years. The first dataset was obtained from the Department for Energy Security and Net Zero (DESNZ), a government organisation which collected fuel productions through surveys of energy suppliers. The dataset is available through the Energy Trends: UK Electricity website in a spreadsheet format [23]. This provided a distinction among different fuel types utilised for electricity consumption along the corresponding month and year. All fuel types available were utilised in the research and these consisted of: Total fuel used, Coal, Oil, Gas, Nuclear, Hydro, Bioenergy, Solar, Low Carbon, Renewables and Fossil fuels. The data spans from January 1995 to October 2023, with over 345 instances, however, it was made public on June 2013. Despite this, there are missing data for certain renewable types: Solar (available from January 2015), Bioenergy (available from January 1997) and Wind (available from January 2007). This implies that recording only occurred when steady production was established.

The weather constituents data are majorly from the Met Office Department, a government department monitoring the UK climate and contributing to associated researchers [24]. This gave us over 100 years of data for temperature (C), precipitations (mm) and average sunshine duration in hours. The data is collected by using grids over the map of the UK, a new reliable method utilised by interpolating from different meteorological stations data onto a uniform grid to provide a standardised average across the UK and take into account different region's area [25, 26]. The dataset are available in a text format (.txt), which required conversion into fixed-width format.

Regrettably, for the wind data, due to missing recording by the Met Office, a separate dataset was required, produced by the DESNZ [27]. In comparison to Met Office's dataset, the wind data is gathered from 12 locations across the UK using station data, and an average is calculated monthly and yearly. Noteworthy is that regions size was taken into account during data collection and appropriate weighting of data points, making as reliable as the gridded data from the Met Office. Conversely, the dataset includes monthly data from 2001 to 2023, thus, average interpolation for the missing data was necessary for this research, specifically for data between 1995 to 2000. The dataset is available as a spreadsheet from the official government website.

In order to calculate per capita electricity consumption, a population dataset was utilised from the Office for National Statistics [28]. This dataset has annually recorded the long-term immigrants, which includes anyone that resides in the UK to live for the period of 12 months. Since 1971, an annual estimate was recorded, however every 10 years, these estimates were corrected by the Census. Moreover, as census is mandatory and protected by UK GDPR and Data protection Act 2018, a full data protection impact assessment was completed [29].

4.1.1 Dataset Licence and Usage

All datasets used during this project are captured under the open government licence for public sector information delivered by the national archives [30].

4.2 Data Cleaning

We meticulously collected six datasets, ensuring their accuracy, and merged the data into a single pandas dataframe. This comprehensive dataset was then saved as a final CSV file for manipulation. The energy dataset, named ET_5.3_JAN_24 containing monthly data points for consumption of individual electricity sources, was downloaded from DESNZ. We collected data points for 12 different sources of electricity across the time frame of 1995 to 2021 including individual sources, total energy consumption, or aggregated totals such as for fossil fuels or sources. Table 1 shows all the variables. The cleaning process involved regrouping these variables and identifying missing data. Decision was made to not modify the energy dataset for wind, solar and bioenergy, this is due to minimal production prior to initial recording year. Furthermore, the government focused more on renewable energy after passing the Climate Change Act 2008, which aims at reducing carbon footprint [31].

Weather	Rainfall, Temperature, Sunshine, Wind
Energy	Total Fuel Used, Coal, Oil, Gas, Nuclear, Hydro, Wind, Bio-energy, Solar, Low Carbon, Renewables, Fossil Fuels
General	Date, Population

Table 1: *Dataset features and targets.*

For the population dataset, the government only provided an accurate yearly dataset. Therefore, a rolling average was used and saved into a dataframe to get a monthly value for the population growth. The monthly average for sunshine, temperature, and rainfall was collected from the online government website for meteorology using the requests package in python [25, 26]. Finally, the last two years from the dataset were set aside to work on it in the future and have them as prediction benchmarks.

Conversely, the wind dataset only includes the monthly data from 2001 to 2023, thus, average interpolation for the missing data was necessary for this research, specifically for data between 1995 to 2000. The dataset is available as a spreadsheet from the official government website.

Before merging the datasets into one, we applied a crucial step: dividing the electricity outputs by the associated population and converting from Mtoe to kWh. This standardisation produces easily intuitive energy data in the form of kWh per capita whilst removing some volatility, as it accounts for population growth with time. The energy sources, dates, and weather patterns were merged into a single dataset called "CompleteDataset" as a CSV file. The combination of weather patterns and energy consumption resources into a singular dataset allows us to perform various manipulations with ease.

4.3 Exploratory Data Analysis (EDA)

4.3.1 EDA with SQL

The initial Exploratory Data Analysis initiated using SQL queries to explore and understand the structure, contents, and relationships within a dataset. Being considered a paramount tool for data analysis and manipulation when working with structured data stored in relational databases.

Followed by this, the SQL queries allow us to prompt commands and functions in order to retrieve, aggregate, filter and perform data manipulation to learn more about the data provided.

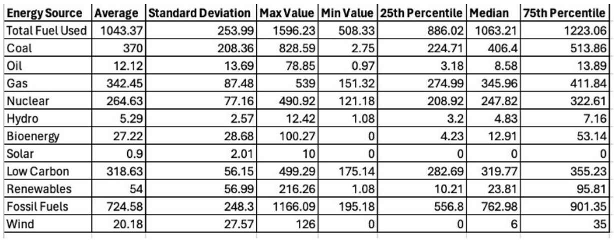
At first, data profiling is used to get an overall idea of the dataset. This shows the basic characteristics such as determining the amount of data points, missing values, number of rows and columns, minimums, maximums and medians when it comes to numerals.

Thereafter, distributions, patterns and relationship within the data have to be investigated. This is done by performing exploratory queries that detects outliers, frequency distribution or even trends. This also allows data cleaning and pre-processing tasks such as how to handle duplicates, missing values or even standardising format.

Finally, performing such EDA using SQL allows quick calculation for computing measures such as kurtosis, skewness and standard deviation.

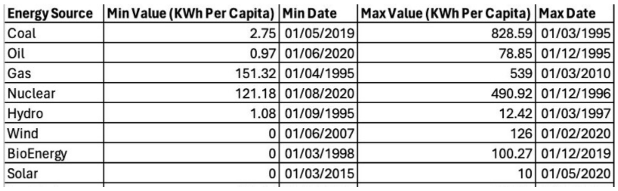
The dataset studied here has been collected from various sources and merge into one. This process has been automated to guarantee no "human" error. There is no missing values in this dataset, 324 data points and 18 columns.

The following tables were generated using SQL and exported on excel for visualisation purposes.



Energy Source	Average	Standard Deviation	Max Value	Min Value	25th Percentile	Median	75th Percentile
Total Fuel Used	1043.37	253.99	1596.23	508.33	886.02	1063.21	1223.06
Coal	370	208.36	828.59	2.75	224.71	406.4	513.86
Oil	12.12	13.69	78.85	0.97	3.18	8.58	13.89
Gas	342.45	87.48	539	151.32	274.99	345.96	411.84
Nuclear	264.63	77.16	490.92	121.18	208.92	247.82	322.61
Hydro	5.29	2.57	12.42	1.08	3.2	4.83	7.16
Bioenergy	27.22	28.68	100.27	0	4.23	12.91	53.14
Solar	0.9	2.01	10	0	0	0	0
Low Carbon	318.63	56.15	499.29	175.14	282.69	319.77	355.23
Renewables	54	56.99	216.26	1.08	10.21	23.81	95.81
Fossil Fuels	724.58	248.3	1166.09	195.18	556.8	762.98	901.35
Wind	20.18	27.57	126	0	0	6	35

Table 2: *Energy sources data description*



Energy Source	Min Value (KWh Per Capita)	Min Date	Max Value (KWh Per Capita)	Max Date
Coal	2.75	01/05/2019	828.59	01/03/1995
Oil	0.97	01/06/2020	78.85	01/12/1995
Gas	151.32	01/04/1995	539	01/03/2010
Nuclear	121.18	01/08/2020	490.92	01/12/1996
Hydro	1.08	01/09/1995	12.42	01/03/1997
Wind	0	01/06/2007	126	01/02/2020
BioEnergy	0	01/03/1998	100.27	01/12/2019
Solar	0	01/03/2015	10	01/05/2020

Table 3: *Energy sources min and max values with associated dates*

During this process, we identified any outliers, missing data, or duplicated rows. Notably, the quartiles of solar energy and wind all indicate 0 (Table 2). This is due to the fact that solar energy was introduced much later (around 2015) and still represents a minimal energy source, averaging 0.9. Furthermore, Table 3 illustrates the shift towards renewable energy.

4.3.2 EDA with Python

Multivariate Analysis examines the relationship between two or more variables concurrently. This was performed in Python, using the following libraries: pandas, numpy, matplotlib and seaborn. Initially, a correlation matrix was displayed between each feature and each target, available in a heat map format (Figure 1). The correlation coefficient chosen was Pearson, due to its popularity and default-value in the pandas library [32]. This analysis demonstrated a negative correlation for temperature and sunshine duration across the majority of features, apart for bioenergy and solar energy sources. The change for solar energy's is the capability to generate more electricity during warmer seasons in turn, increasing solar energy consumption. Whereas, bioenergy should not be influenced by weather patterns, due to its nature

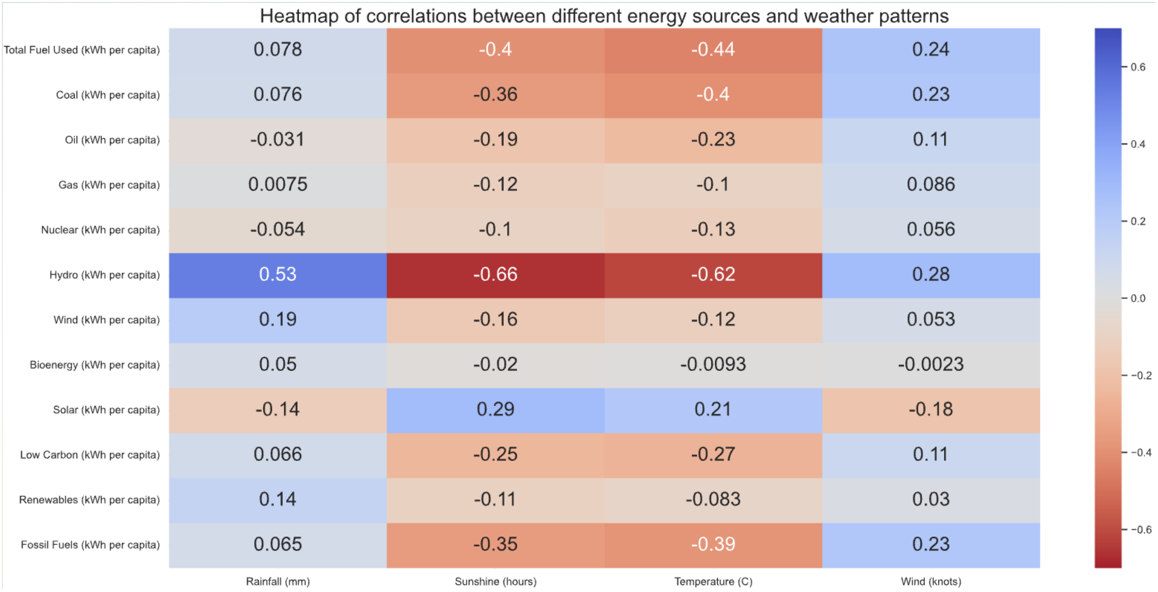


Figure 1: *Heatmap of correlations between different energy sources and weather patterns*

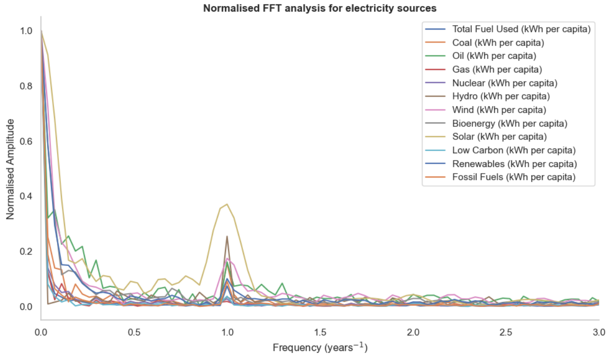


Figure 2: *Fast Fourier transformer of consumption of electricity sources*

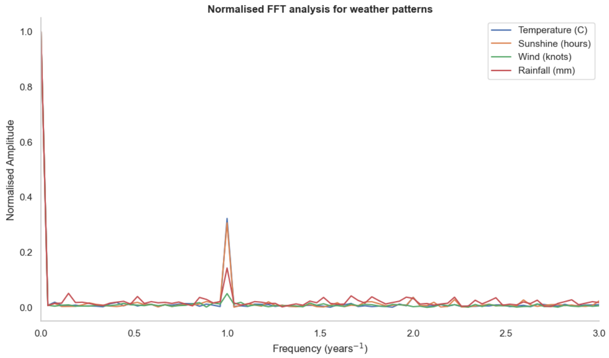


Figure 3: *Fast Fourier transformer of weather patterns*

[33]. Notably, hydro energy has a strong negative coefficients between both temperature and sunshine duration, suggesting a lower energy consumption during summer. There is a positive correlation between rainfall and wind for many electricity sources. Hydro energy has a significantly strong positive correlation with rainfall, due to its operational dynamics.

Thereafter, a Fast Fourier Transformer (FFT) analysis was conducted on all features and targets, in order to find the presence frequency based patterns. The FFT produces a spectrum containing sequences of components that form the signal as a sum of sine waves using the following formula:

$$x[k] = \sum_{k=0}^{N-1} x[n] e^{-j\frac{2\pi}{N}kn} \tag{1}$$

Where x[n] is the input sequence. In this case, the FFT computes efficiently the Discrete Fourier Transform and its inverse, by using the time complexity of $O(n \log n)$ rather than $O(n^2)$, where n is the number of points. As you can see from Figure 2 and Figure 3, the mutual shared fundamental frequency, of per year, implied a repeating seasonal pattern for features and targets. This is further supported by exploring the time-series graphs of energy sources and weather patterns (found in Appendix B), proving the alternating values between summer and winter.

Overall, the EDA conducted with python comprehensively explored the relationships between targets and features, discovering shared seasonal patterns in terms of electricity consumption by source and weather features.

4.4 Predictive Modelling

All models underwent uniform application across identical randomised training and testing sets, with the training set comprising 80% of the data and the testing set 20%, to ensure fair comparison. The train-test split was created using the sklearn.model_selection.train_test_split function. Feature scaling was executed using sklearn's StandardScaler function. This was done due to the large variation in variable sizes between features ensuring large valued features did not skew the models. The target variables also had a large disparity in scale spanning several orders of magnitude for example the mean values for solar and coal energy production was 0.9 and 370 kWh per capita respectively as seen in Table 2. Therefore target variables were also scaled to allow fair comparison of model performance metrics between targets.

After training and testing, the predicted values were evaluated against ground truth values outputting several performance metrics and graphically displaying the predicted against the ground truth to allow intuitive analysis. After careful consideration of performance metrics; models were evaluated with the widely used R^2 and MSE performance metrics following the equations:

$$R^2 = 1 - \frac{\text{Sum of squares of residuals}}{\text{Total sum of squares}} \tag{2}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{3}$$

Where n is the number of testing data points, Y is the predicted value and $\hat{Y}_i$ is the ground truth value.

R^2 values range from -1 to 1. With $R^2 = 1$ being a perfect fit; $0 < R^2 < 1$ meaning the model explains some but not all of the variation in the problem with higher values indicating better fit; $R^2 = 0$ meaning there is no relationship between the features and target explained by the model; $R^2 < 0$ implies an extremely poor fit. The MSE is a positive value that quantifies the average accuracy of the model with lower MSE values implying better fit. This allowed a comprehensive understanding of precision and accuracy of model performance allowing intuitive portrayal of model effectiveness. This standardized approach enabled an objective assessment of each model's effectiveness, aiding in the identification of the most suitable model for the given task.

The use of cross validation across all models was attempted in order to provide a more reliable model performance whilst reducing the possibility of overfitting. The use of cross validation would also maximise efficient use of data used in both testing and training however, due to the relatively small dataset, this negatively affected several model's effectiveness.

This combined with the large increase in computational cost of models, especially the neural network and random forest regression models, meant that the use cross validation could not be effectively applied to all models used therefore reducing the reliability of model performance comparison. Consequently, the use of cross validation was discarded.

4.4.1 Multiple Linear Regression

Linear regression (LR) is the simplest regression model used in many tangential research papers [34, 35, 36, 37].

Multiple LR is an extension of standard LR including multiple features each linearly regressed onto one target as seen in the equation below.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n \tag{4}$$

Where Y is the target, X is the feature, β is the coefficient of the corresponding features.

Several algorithms are available for multiple LR. After testing a selection of simple LR models available within Sklearn's library, the linear support vector regression (LinearSVR) model was chosen in this project. Support vector regression (SVR) was chosen due to its effectiveness in high dimensional space, memory efficiency, versatility as well as superior result in comparison to LR models trialed. Sklearn's SVR was used with a linear kernel, regularisation parameter of 1 and epsilon of 0.1. LinearSVR uses kernel function to map the features into a high-dimensional space in order to find a linear relationship to the target. The relationship between features and the target is approximated to a linear relationship in this high-dimension space. Support vector regression uses a margin of tolerance in order to minimise the error within a set margin; penalising larger errors outside it. SVR models aim to minimise the loss function whilst maximising the margin of tolerance. This is the distance between the hyperplane and the closest data points also known as the support vectors.

4.4.2 Neural Networks

Neural networks (NN) are a versatile and flexible method of machine learning that can model complex non-linear relationships between input and output variables. If given sufficient data they allow effective modelling of complex problems with hidden patterns. NN were seen in several papers of a similar problem to be the most effective model tested [6, 8, 7]. The TensorFlow library was used for the creation of the NN due to its ease of use and well explained documentation. The architecture of the NN used consisted of three layers and was defined by the nature of the problem.

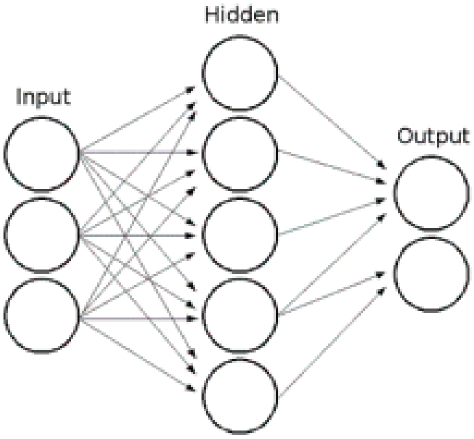


Figure 4: *Schematic diagram of the generalised architecture of a three-layer neural network.*

The NN created for this project used optimiser Adam; an input shape of four (the number of input features); the first and second layer consisted of the TensorFlow default of sixty four neurons; and an output layer of one neuron. Optimiser Adam is a stochastic gradient descent method and was chosen due to its applicability to the dataset size as well as suitability for noisy data. Both the first layer acted both as an input layer and a hidden layer receiving the input data whilst being used to perform computation. The second layer acted purely as a hidden layer and was only used for computation processing the representations passed through by the first layer.

Both the first and second layer performed nonlinear transformations on the received data. The output layer aggregated information produced by the previous layers and produced a prediction for the target variable.

One challenge encountered while employing NN was their susceptibility to over-fit. To mitigate this issue, adjustments were made to the number of neurons in the first two layers. These alterations yielded minimal impact on the performance and so the neuron number was returned to the original values. Subsequently, the use of dropout regularisation showed promising signs in reducing over-fitting and so was incorporated. The dropout rate was set to 0.3 and was applied to the first two layers. This allowed 30% of the neurons in the associated layers to be randomly dropped during training. Dropout is not applied during testing however outputs from neurons are scaled by the dropout rate. Dropout regularisation allows the development of a more generalised and robust model reducing over-fitting.

4.4.3 Extreme Gradient Boosting (XGBoost)

Released in 2014, XGBoost library was considered the go to machine learning program during hackatons for its speed and efficiency. Belonging to the ensemble learning, it uses decision trees as base learners before applying a regularisation technique to enhance the model. Moreover, XGBoost performs very well while dealing with missing values and can be applied for a variety of tasks like classification or ranking.

In this research, the implementation called on the XGRegressor function, using weather sources as features and energy sources as targets. In order to improve the output a hyper parameters tuning functions was created using GridSearchCV and find the most optimal parameters to avoid over fitting as this is one of the limitation of XGBoost [39]. Doing so made the program run much faster once the optimal hyper parameters were determined. It is important to note that XGBoost has exceptional performance in a variety of machine learning tasks and has built-in regularisation techniques such as L1 or L2. Finally, to take it one step further, a "month lag" has been programmed, this is to make the algorithm use n previous months of data to perform calculations.

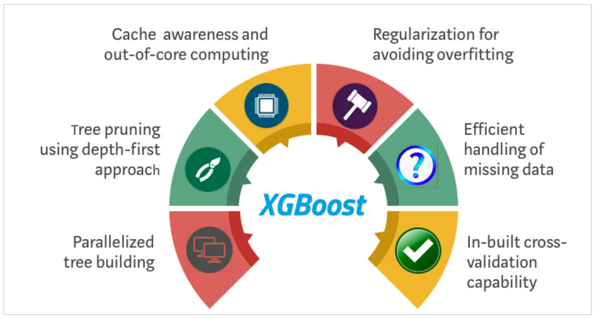

Cache awareness and out-of-core computing
Regularization for avoiding overfitting
Tree pruning using depth-first approach
Efficient handling of missing data
XGBoost
Parallelized tree building
In-built cross-validation capability

Figure 5: *Steps of XGBoost optimising the standard GBM algorithm [38].*

Finally , despite its effectiveness, XGBoost is susceptible to over-fitting. It necessitates careful tuning of hyper parameters such as tree depth, learning rate, and regularization parameters to counteract over-fitting tendencies and enhance generalization performance. Additionally, XGBoost operates under the assumption of an additive and linear relationship between features and the target variable, which may not always accurately reflect real-world scenarios.

4.4.4 Random Forest Regression

Random forest regressor (RFR) from Sklearn, is a machine learning algorithm using an ensemble of decision trees that makes a prediction based on the average of all trees. This technique involves building multiple decision trees, each containing a sample of the data using the bootstrap bagging method. Figure 6 below illustrates the model development of RFR. The trees in the forest use the a best split strategy, the sub sampling can be controlled with the max_samples parameter and the bootstrap parameter was set to true to ensure trees were sub sampled with random parts from the dataset, otherwise the whole dataset is used.

It is commonly used as a classification, regression, or clustering tool to make accurate predictions. The nature of the random forest results in higher accuracy over the data since it is less prone to over fitting since the ensemble of tree decision predictions is averaged. The electricity consumption and weather pat-

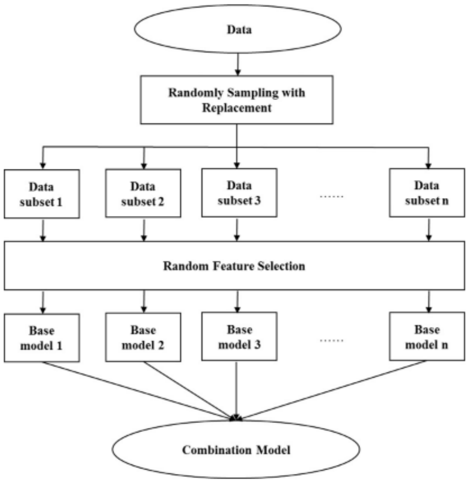


Figure 6: *Diagrammatic representation of a random forest model [40].*

terns datasets usually have unusual patterns and outliers due to spikes in energy consumption. As a quick example, the UK saw a spike in energy consumption in late February of 2018, as a cold snap called the "Beast from the East" severely impacted home and road infrastructure in the country. The random forest is robust against noise and outliers in the data meaning it is able to effectively handle such patterns without significantly impacting its performance and results. We are also trying to study feature importance, which makes this method very useful as it assigns various weights to different features, using the gini coefficient while calculating the output of each tree.

5 Results

5.1 Predictive modelling

5.1.1 Multiple Linear Regression

The multiple LR model was found to be too simple to capture the complexity of the problem leading to poor predictions. Figure 7 shows that predictions produced by the LinearSVR model ignores patterns in the data returning an almost constant predicted value of the target. This implies that there are hidden patterns within the data that cannot be captured through a simple multiple LR model.

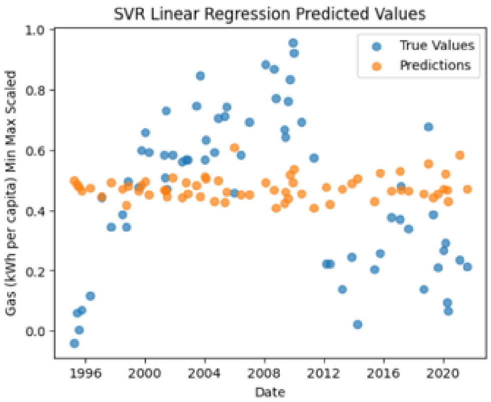


Figure 7: *Example SVR linear regression for predictions and true values for gas showing poor fit.*

5.1.2 Neural Network

The NN model captured the complexity of the problem except also lead to over-fitting. The over-fitting can be seen in figure 8 by the increase of testing error with respect to the training error with number of epoch.

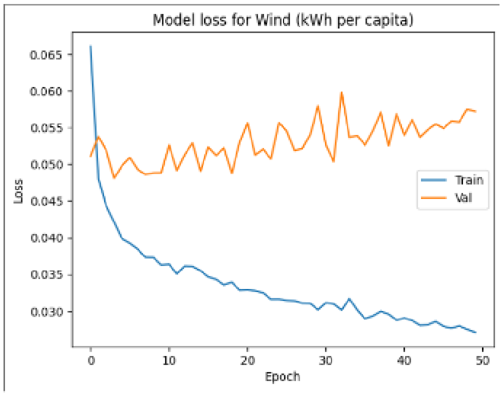


Figure 8: *Training and validation set errors against number of epochs for wind energy for the neural network without regularisation.*

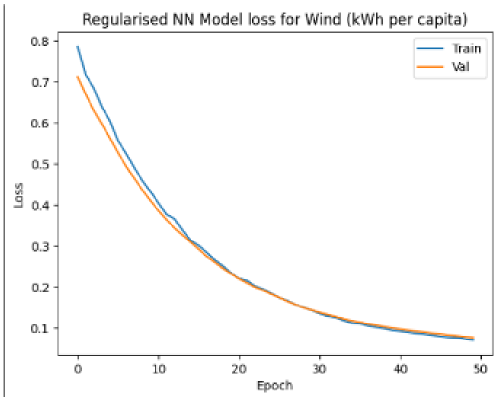


Figure 9: *Training and validation set errors against number of epochs for wind energy for the neural network with dropout regularisation.*

After dropout regularisation was applied to the NN model the effect of over fitting was reduced. As seen in figure 9 showing that the training and testing errors converge with epoch for the same model applied to the same target using the same test-train split.

NN were found to have a large variation in performance metrics between runs giving different values of R^2 and MSE. This implies that the NN model is very susceptible to the effects of outliers that can be caught within different testing sets during the randomised train-test split.

5.1.3 Extreme Gradient Boosting (XGBoost)

After rewriting appropriate code for the dataset studied XGBoost performances shows promising results between hydro energy and the electricity consumption per capita. The output between ground truth and predicted values seems to show the model's effectiveness.

Adding a lag of n months makes (see code portion in Appendix A) XGBoost perform more efficiently. Doing so allows the model to capture temporal dependencies in the data as well as providing more information about past trends and patterns. To summarise, adding a lag increases the model efficiency and leads to more accurate predictions in time series forecasting tasks. It is clear that Figure 11 shows better results compare to the XGBoost without lag seen in Figure 10.

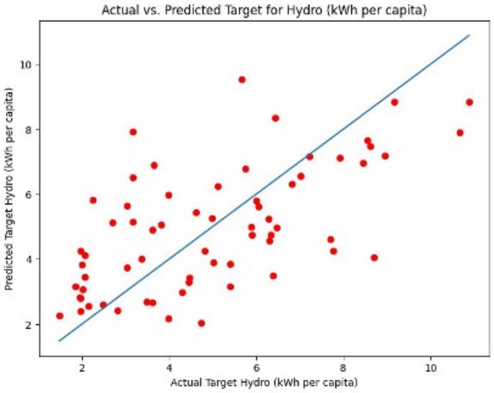


Figure 10: *XGBoost predictions against true for hydro energy.*

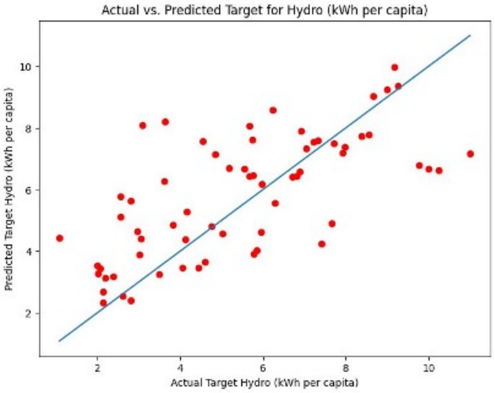


Figure 11: *XGBoost with 5 month lag predictions against true for hydro energy.*

5.1.4 Random Forest Regression

RFR emerges as the most effective model for our dataset. Its performance, robustness in handling missing data and outliers, and ability to provide reasonably accurate predictions are evident. However, it's important to emphasize the need to note the nuanced outcomes observed in different industry sectors.

Our analysis, as depicted in Figure 12, underscores the promising potential of RFR in the hydro industry. Despite the predicted points not aligning perfectly with the ideal trend line, a discernible trend indicates fair predictions over time.

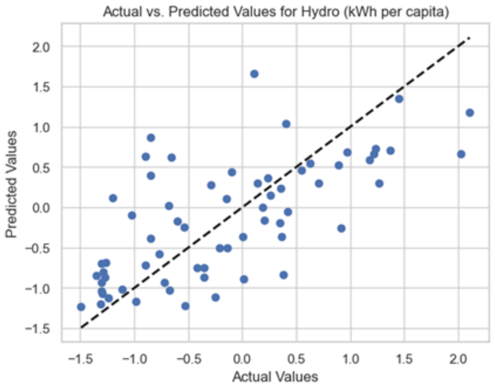


Figure 12: *Random forest regression predictions against true for hydro energy.*

Conversely, Figure 13 showcases predictions for the gas industry. The absence of a clear trend line suggests a nuanced relationship between weather patterns and energy consumption.

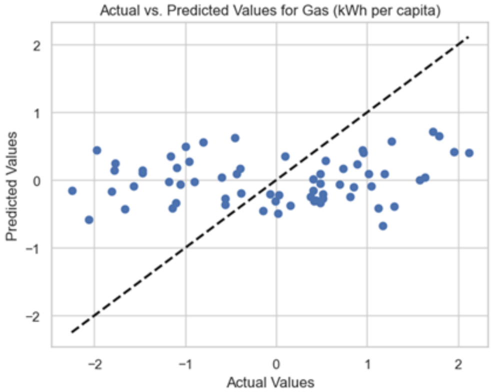


Figure 13: *Random forest regression predictions against true for gas.*

Similar to NN, RFR imposes significant computational demands and high memory usage. This is due to the need for data replication for each decision tree (over a 100 times). The computational intensity of RFR is a factor that requires careful consideration, especially in large-scale applications.

5.2 Model Evaluation

5.2.1 R^2 Metric

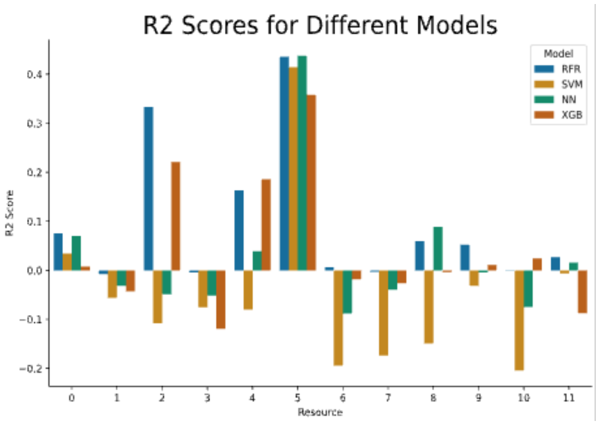


Figure 14: *R2 Performance metric evaluation of models. Resources are indexed as total fuel$^{(0)}$, coal$^{(1)}$, oil$^{(2)}$, gas$^{(3)}$, nuclear$^{(4)}$, hydro$^{(5)}$, wind$^{(6)}$, bioenergy$^{(7)}$, solar$^{(8)}$, low carbon$^{(9)}$, renewables$^{(10)}$ and fossil fuels$^{(11)}$.*

As seen in figure 14 there was a large variation in R^2 values between models and targets. Although none of the models perfectly modelled the problem, the RFR model was consistently the most effective across multiple targets. LinearSVM had the poorest performance across multiple targets frequently giving negative R^2 values. All models for coal, gas and bioenergy returned negative values implying that the models were very ineffective for these targets. Hydro energy was the most effectively modelled target with highest R^2 values of all targets with NN giving the highest R^2 value of 0.431. RFR showed promise for effective modelling of oil, nuclear and hydro with XGB having comparible results for these targets.

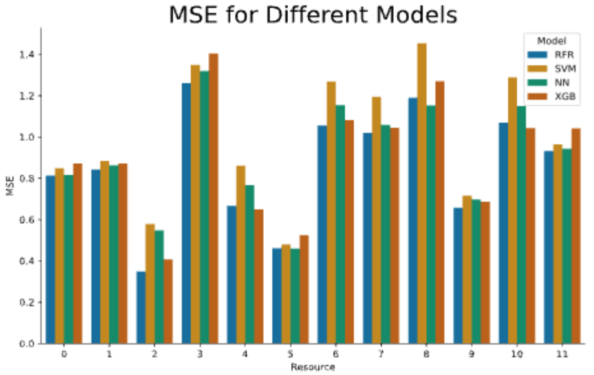


Figure 15: *Scaled MSE Performance metric evaluation of models. Resources are indexed as total fuel*$^{(0)}$, *coal*$^{(1)}$, *oil*$^{(2)}$, *gas*$^{(3)}$, *nuclear*$^{(4)}$, *hydro*$^{(5)}$, *wind*$^{(6)}$, *bioenergy*$^{(7)}$, *solar*$^{(8)}$, *low carbon*$^{(9)}$, *renewables*$^{(10)}$ *and fossil fuels*$^{(11)}$.

5.2.2 Mean Square Error

Figure 15 shows the scaled MSE values for each model for each target. Target variables had to be scaled due to the large disparity in magnitude causing the MSE to be scaled which allows comparison between targets. The difference in MSE between models for individual targets is less apparent than the difference in R^2 values however it can be seen that the RFR model consistently produced the lowest MSE values throughout with only one exception for Nuclear energy. LinearSVM frequently gave the poorest MSE values. XGBoost in comparison gave very similar results to the RFR model for many targets but also gave several very poor MSE values in comparison to the other models for specific targets.

Hydro and oil electricity sources returned the lowest MSE values across models with RFR model giving the lowest MSE value of the project when predicting consumption of electricity produced by oil.

6 Discussion

This study explored the relationship between weather conditions and electricity consumption by source. Effects on the electricity consumption due to seasonal changes can be seen as all variables shared a fundamental frequency of once per year found using Fourier transform. Correlation coefficients found during the EDA implied linear relationships between individual features and several targets, however many correlations were extremely weak implying either no relationship or that the relationship is not linear. Hydro electricity showed the strongest linear relationships with respect to all features. This implies a potential for modelling using multiple Linear Regressions which was validated by the large R^2 and low MSE values for the LinearSVR model.

Taking into account both the R^2 and MSE metrics the performance of each model for each target can be assessed. Although none of the models used showed perfect fit to any of the targets, RFR returned the most promising and consistent results across both performance metrics. LinearSVR consistently performed the worst for all targets across both performance metrics. It is clear that the problem cannot be effectively modelled using multiple linear regression as there are hidden patterns in the data that cannot be captured using purely linear models. NN had middling performance in comparison to the other algorithms trialed. Although having the best overall performance for hydro electricity across both metrics, the NN performance was limited and returned very poor results for all other targets. The NN model was limited due to its susceptibility to over-fit and sensitivity to outliers. XGBoost had similar results to RFR with negligible difference between model performance for nuclear electricity production prediction. It performed well on oil, nuclear and hydro electricity sources. XGBoost's assumption of linearity between feature and target is one of its limitation that causes such difference with RFR. The RFR model shows good promise for the ability of modelling electricity consumption by source from weather features. RFR can capture the non-linear relationships between the features and targets studied. It is very likely in this scenario that weather features have a non-linear effect on the electricity consumption. Moreover, RFR returned good results in performances metrics. This makes RFR the most suitable choice for modelling.

Although weather features cannot capture all variance in electricity consumption, it is clear that weather features are an important factor in the creation of a successful model making a large impact on energy consumption depending source. This report proves that our dataset is extensive enough for forecasting certain target electricity sources but lacks influential variables.

Since renewable energy sources are dependant on weather conditions, it was expected that these targets would have the most effective models. However, there has been a rapid growth in the production capacity of many renewable electricity sources leading to a clear overall trend in data with respect to time that cannot be captured effectively with weather feature variation. Hydro electricity production capacity undergoes negligible change through the time range of the project and so is not affected by this trend. This somewhat explains why the hydro electricity had the best model fits of all targets irrespective of which algorithm was used.

Performance on gas returned high MSE values and negative R^2 values for all models showing that gas electricity consumption could not be modelled effectively implying no relationship with weather conditions. This inability to model gas could also stem from the mass movement of manufacturing out of the UK significantly reducing per capita energy consumption which, as seen in the SQL data exploration is the main source of electricity consumption in the UK. This overall trend in the reduction of electricity consumption per capita can be seen in Appendix B as Figure 16.

Due to the late implementation of several renewable sources of electricity, such as bioenergy, solar energy and wind energy, a limitation of our project is the prevalence of zero values in these sources. This can be seen in poor correlation coefficients for these targets. Perhaps having a full dataset would have allowed more effective modelling and in turn given better results for the MSE and R^2, as shown with hydro. Additionally, another drawback is the sample frequency of our datasets, as they are all monthly. Therefore, using a higher sample frequency, like daily data, could give a higher resolution view of the patterns, allowing the determination of small time frame perturbations in electricity consumption caused by variation in weather features therefore enhancing model performance.

Potential for future improvements include the application of cross validation across all models learning to more robust models with more reliable performances. This would reduce the effect of outliers which greatly affected the NN model. However, as discussed above the application of cross validation was not possible for the case of this project. Parameter tuning played a substantial role in this project and fine tuning the parameters further would lead to significantly more effective models. Testing the models on an isolated testing set would allow more reliable evaluation of models however due to the number of data points using an isolated testing sample would have significant negative effects on the robustness of the models used. An investigation could be done into which weather feature has the greatest importance in modelling electricity consumption. Inclusion of high impact variables such as importing and exporting, economic influence and major shifts in manufacturing.

7 Conclusion

This paper focused on modelling the energy consumption per capita by using different weather patterns, as supported by the correlations coefficient found in the visualised EDA. The usage of various models allowed to ensure fair comparisons of the performance on the dataset using MSE and R^2 metrics. RFR was consistently the best performing model for this dataset, with predictions for hydro being most accurate, also supported by the high correlation in the EDA, showcasing the impact of weather on different energy sources. Unexpectedly, oil and nuclear were also found impacted by the weather condition. It is anticipated that due to reliability these energy sources are used to compensate for the overall increase in energy requirements that cannot be offset by other sources leading to this trend.

Overall, this research fills an existing gap in the literature by exploring the affect of different electricity sources against various weather constituents and which models are best suited to make adequate predictions. This research could benefit from additional refinement through the data points availability, particularly for majority of the renewable energy types. We proved that modeling different electricity sources was vital in the UK, as no one else has studied the fragmented market this way. Weather, although a small factor in energy consumption, has been proven in this paper to have major impacts on hydro, nuclear and oil.

References

[1] F. Kaytez, M. C. Taplamacioglu, E. Cam, and F. Hardalac. Forecasting electricity consumption: A comparison of regression analysis, neural networks and least squares support vector machines. *International Journal of Electrical Power & Energy Systems*, 67:431–438, 2015.

[2] H. Son and C. Kim. Short-term forecasting of electricity demand for the residential sector using weather and social variables. *Resources, Conservation and Recycling*, 123:200–207, 2017.

[3] T. Zhang, X. Zhang, O. Rubasinghe, Y. Liu, Y. Chow, H. Iu, and T. Fernando. Long-term energy and peak power demand forecasting based on sequential-xgboost. *IEEE Transactions on Power Systems*, page 1–16, 2023.

[4] A. Abbasi and A. Norouzi. Short term load forecasting using xgboost. *Springer*, 2019.

[5] B. Dhupia. Ensemble machine learning modelling for medium to long term energy consumption forecasting. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(10):459–463, 2021.

[6] T. Yang, Y. Chen, J. Emer, and V. Sze. A method to estimate the energy consumption of deep neural networks. *Massachusetts Institute of Technology*, 2017.

[7] N. Jaisumroum and J. Teeravaraprug. Forecasting uncertainty of thailand's electricity consumption compare with using artificial neural network and multiple linear regression methods. *Department of Industrial Engineering Faculty of Engineering, Thammasat University Pathumthani*, 2017.

[8] G. K. F. Tso and K. K. W. Yau. Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. *Energy*, 32(9):1761–1768, 2007.

[9] X. Pang, C. Luan, L. Liu, W. Liu, and Y. Zhu. Data-driven random forest forecasting method of monthly electricity consumption. *Electrical Engineering*, 104(4):2045–2059, 2022.

[10] M. Kitson and J. Michie. *The Deindustrial Revolution: The Rise and Fall of the UK Manufacturing 1870-2010*. 2014. Available from: https://www.jbs.cam.ac.uk/wp-content/uploads/2023/05/cbrwp459.pdf, [Accessed: 28-02-2024].

[11] A. Cangiano. The impact of migration on uk population growth - migration observatory, 2023. Available from: https://migrationobservatory.ox.ac.uk/resources/briefings/the-impact-of-migration-on-uk-population-growth/, [Accessed: 05-03-2024].

[12] C. L. Hor, S. J. Watson, and S. Majithia. Analyzing the impact of weather variables on monthly electricity demand. *IEEE Transactions on Power Systems*, 20(4):2078–2085, 2005.

[13] A. Prabakar, L. Wu, L. Zwanepol, N. Velzen, and D. Djairam. Applying machine learning to study the relationship between electricity consumption and weather variables using open data. 2018.

[14] H. Saima, J. Jaafar, S. Belhaouari, and T. A. Jillani. Intelligent methods for weather forecasting: A review. *2011 National Postgraduate Conference*, 2011.

[15] A. D. Amato, M. P. Ruth, and J. Horwitz. Regional energy demand responses to climate change: Methodology and application to the commonwealth of massachusetts. *Climatic Change*, 71(1-2):175–201, 2005.

[16] P. M. Maçaira, R. C. Souza, and F. L. C. Oliveira. Forecasting brazil's electricity consumption with pegels exponential smoothing techniques. *IEEE Latin America Transactions*, 14(3), 2016.

[17] E. Ceperic, V. Ceperic, and A. Baric. A strategy for short-term load forecasting by support vector regression machines. *IEEE Transactions on Power Systems*, 28(4):4356–4364, 2013.

[18] C. Luan, X. Pang, Y. Wang, L. Liu, and S. You. Comprehensive forecasting method of monthly electricity consumption based on time series decomposition and regression analysis. *DTU*, 2020.

[19] J. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandawalle. Least squares support vector machines. *Singapore: World Scientific Publishing*, 2002.

[20] Office for National Statistics. Manufacturing and production industry, 2024. Available from: https://www.ons.gov.uk/businessindustry and-trade/manufacturingandproductionindustry, [Accessed: 03-03-2024].

[21] Office for National Statistics. Energy consumption in the uk (ecuk) 1970 to 2019, 2020.

[22] J. Hackney. Increasing the value of weather information in the operation of the electric power system technical report. *Environmental and Societal Impacts Group (ESIG)*, 171, 2002.

[23] UK Gov. Energy trends: Uk electricity, 2024. Available from: https://www.gov.uk/government/statistics/electricity-section-5-energy-trends,[Accessed: 18-02-2024].

[24] Met Office. Uk and regional series, 2019. Available from: https://www.metoffice.gov.uk/research/climate/maps-and-data/uk-and-regional-series, [Accessed: 18-02-2024].

[25] D. Hollis, M. McCarthy, M. Kendon, T. Legg, and I. Simpson. Haduk-grid—a new uk dataset of gridded climate observations. *Geoscience Data Journal*, 6(2):151–159, 2019.

[26] J. D. Ramsdale, M. R. Balme, S. J. Conway, C. Gallagher, S. A. van Gasselt, E. Hauber, C. Orgel, A. Séjourné, J. A. Skinner, F. Costard, A. Johnsson, A. Losiak, D. Reiss, Z. M. Swirad, A. Kereszturi, I. B. Smith, and T. Platz. Grid-based mapping: A method for rapidly determining the spatial distributions of small features over very large areas. *Planetary and Space Science*, 140:49–61, 2017.

[27] UK Gov. Energy trends: Uk weather, 2024. Available from: https://www.gov.uk/government/statistics/energy-trends-section-7-weather, [Accessed: 18-02-2024].

[28] Office for National Statistics. Estimates of the population for the uk, england and wales, scotland and northern ireland - office for national statistics, 2022. Available from: https://www.ons.gov.uk/peoplepopulation andcommunity/populationandmigration /populationestimates/datasets/ populationestimatesforukengland andwalesscotlandandnorthernireland, [Accessed: 18-02-2024].

[29] ONS. Protecting personal data in census 2021 results, 2023. Available from: https://www.ons.gov.uk/peoplepopulation andcommunity/populationandmigration/ populationestimates/methodologies/ protectingpersonaldataincensus2021results, [Accessed: 18-02-2024].

[30] The National Archives. Open government licence, 2019. Available from: https://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/, [Accessed: 18-02-2024].

[31] B. Hilary. Climate change act 2008, 2011. Available from: https://bills.parliament.uk/bills/195, [Accessed: 05-03-2024].

[32] P. Schober, C. Boer, and L. A. Schwarte. Correlation coefficients: Appropriate use and interpretation. *Anesthesia & Analgesia*, 126(5):1763–1768, 2018.

[33] P. E. Allen and G. P. Hammond. Bioenergy utilization for a low carbon future in the uk: the evaluation of some alternative scenarios and projections. *BMC Energy*, 1(1), 2019.

[34] M. Maaouane, S. Zouggar, G. Krajačić, and H. Zahboune. Modelling industry energy demand using multiple linear regression analysis based on consumed quantity of goods. *Energy*, 225:120270, 2021.

[35] N. Fumo and Rafe Biswas. Regression analysis for prediction of residential energy consumption. *Renewable and Sustainable Energy Reviews*, 47(47):332–343, 2015.

[36] E. Kafazi, R. Bannari, A. Abouabdellah, My. Aboutafail, and J. Guerrero. Energy production: A comparison of forecasting methods using polynomial curve fitting and linear regression. In *2017 International Renewable and Sustainable Energy Conference (IRSEC)*, pages 1–5, 2017.

[37] V. Bianco, O. Manca, and S. Nardini. Linear regression models to forecast electricity consumption in italy. *Energy Sources, Part B: Economics, Planning, and Policy*, 8(1):86–93, 2013.

[38] V. Morde. Xgboost algorithm: Long may she reign! 2019. Available from: https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d, [Accessed: 18-02-2024].

[39] xgboost developers. Xgboost documentation — xgboost 1.5.1 documentation, 2022. Available from: https://xgboost.readthedocs.io/en/stable/, [Accessed: 18-02-2024].

[40] Zeyu Wang, Yueren Wang, Ruochen Zeng, Ravi S. Srinivasan, and Sherry Ahrentzen. Random forest based hourly building energy prediction. *Energy and Buildings*, 171:11–25, Jul 2018.

7.1 Group Work

From start to finish, us, the triangles have always been very well coordinated. From the choice of general topic we have brainstormed ideas, weighted pros and cons, likes and dislikes, but most importantly always listen to each others opinions. Collaborating on the notion software we have managed to organise every piece of information collected whether is code, datasets, PowerPoint, checklists or meeting monitoring.The reason why we could also perform well is because of our regular meetings on campus but also on Microsoft teams if a team member is not available.

For the duration of this project we have met thirty times that resulted in about fifty hours. As a team we also had disagreements which were solved in a timely and calmly manner by bouncing ideas between us and discussing everything openly. Our easy communication between us might be assisted by the fact that we were all friends from the start and not just colleagues.

This project was quite extensive and therefore an equal division of the tasks was paramount for efficiency but also fairness within the group. All of us have performed an equal amount of work and have been involved in all the aspects of the project from the data collection all the way to the result analysis.Throughout our presentations we displayed a good team work in terms of speaking time. Moreover, we all understood each others sections which made the writing of this report go smoothly however it was still time consuming. As we were all well-acquainted with the entire project, we could distribute the workload among ourselves in a an equal manner and overcome this final step before submitting our research.

Finally, we would like to thank our supervisors, Dr Qamar Natsheh and Nizar Al Ahmad for their great help and guidance. The parts they played in our group have been very helpful.

8 Individual Contributions

8.1 Ali Abouyahia

Working with Sam, Mohamad and Aman has been an amazing experience from start to finish.The choice behind weather and energy was a collaborative decision as it was something we all wanted to investigate.It was also a great choice in terms of data availability as the weather data is usually something public. With my experience I already had an idea on how we could tackle the problem as I have done a similar research during my previous masters and during my work experience.After establishing the question and divided all tasks among us I started right away collecting data from various government sources and querying on said data to perform "pre" Exploratory data Analysis (EDA).However, a requirement we were not aware of was to come up with a single dataset and not several tables. Once the complete dataset was created I was in charge of performing the EDA in SQL to look for outliers , duplicated rows or any missing data.An observation was also done on the quartiles and showed relevant information regarding the evolution of renewable energy sources through time.Moreover, while still using SQL I have made separates tables using sub-queries that summaries in a more concise manner extra information such as the percentage of energy used per year , the highest and lowest energy source with their corresponding dates. To establish a correlation between weather and energy we turned to machine learning and selected the following four techniques : XGBoost ,Neural Networks,Multiple Linear Regression and Random Forest. I, myself, worked on XGBoost as I was already familiar with.Using the XGRegressor function in python and by tweaking hyper parameters I came up with two extensive codes.On one hand , a "basic" application of XGBoost on our data which returned "normal" results quite efficiently.On the other hand , I have added a function that made the algorithm select more previous data to make its calculation and then the results were much better but slightly slower.This part was challenging as finding the correct hyper parameters can be very time consuming whether it is done manually or by coding. This project is not only about computer science.It is also about communication.There has been four submissions we were all in charge of submitting one , in my case it was our first presentation which I also proof read. Working closely with Sam I came up with the layout of our final presentation as well as the contents for the slides in EDA in SQL and XGBoost.Moreover, to make sure that we were on track and not forget anything I have also kept track of our meeting times and notes via Notion. In order to group all our work in a nice and professional way I came up with the idea to use GitHub, share it among all of us to have access for upload.That way we all get to contribute and not only the creator of the repository (Sam) has to upload everything on his own.Regarding the report I wrote the sections I implemented which were EDA with SQL, XGBoost methodology, results section for XGBoost the abstract and finished by proofreading the full report before submission. Finally , working as a group has been very smooth and would like to thank my classmates for this experience and hope to work with them in the future on other projects.

8.2 Amanjot Singh

This group project has been a great experience for sharing and applying knowledge gained throughout this Master program. Thanks to the open communication with the rest of the team members, Ali, Mohamad and Sam, we have been able to enjoy the process while passing through different stages. Since the initial day, we decided the topic being in the energy sector and weather, therefore, the main objective was to find reliable data. The datasets for all different weather constituent and population data were collected by me. This involved the exploration and assessing various dataset's reliability for this topic. A challenge faced during the collection, was the conversion from .txt file to fixed-width format. The goal was to clean the data using only python; therefore, an automated algorithm was made to download the data from the website itself, and then save it as a fixed-width format which allowed to open it as a data frame with the pandas module. Although with this challenge, now our code can be run on all computer without requiring a specific path. Also, I cleaned the different weather datasets, apart for the wind, and made initial visualisation for the progress check. Afterwards, my focus was on perfecting the exploratory Data Analysis, as it allows to make examination and visualisation of the data frame in python. This allowed the rest of the team, to understand key patterns and characteristics of our research. Besides the fact that majority of the EDA could have been done on python; however, we utilised SQL to query specific key trends to showcase our diverse programming knowledge. This took additional time than initially planned, as I was making some background research to support the EDA findings, and make sure the rest of the team understood them prior to the machine learning. Subsequently, then my contribution was working on the Neural Network, specifically on adding the regularisation to the algorithms and run it. This was important as it allowed to minimise over-fitting of the model, and therefore get optimal results for it. In addition, we all participated on the creation of the progress check, poster and presentations together. I was involved creating the template and fill the text with the consultation of my team. In the creation of this report, I have written the introduction, data collection, the EDA using Python, Conclusion and helped Sam with the Discussion. Additionally, I have proof-read the report and formatted the figures. Overall, my contribution has been on finding suitable datasets, undertake the first exploratory data analysis, and then help out with the neural networks. Overall, I have enjoyed working in this group project, and it has allowed to show my skills in a practical and interesting project.

8.3 Mohamad Abdallah

Contributing with Sam , Ali and Aman has been fantastic from the start we were almost always on the same page and clicked right away.Throughout our project journey, I contributed significantly to various aspects aimed at understanding the interplay between weather patterns and the energy sector. Early on, I collaborated with the team to shape our research question, ensuring it aligned with our goals. I was also involved in researching the various datasets required for our project. To keep us organized, I set up a system on Notion, a note-taking application for easy access to files, project timelines, and sample codes for our project while also building a Gantt chart to help us stay on track. Streamlining our project workflow, from data import to machine learning and dashboard creation, was also part of my role, ensuring each step flowed smoothly. Data management was a crucial aspect of our project, and I took the lead in this area. I was responsible for importing energy datasets and cleaning them for analysis. I also gathered similar work my team members did for the other datasets and grouped them into a singular dataframe. This meticulous process ensured the final dataset was accurate and reliable, ready for general use. As part of our predictive (ML) analysis, I implemented a Random Forest Regression and minorly tuned it, which provided some of our best results. After collecting the Neural Network and Support Vector Machine from Sam and XGBoost from Ali, I graphed our results onto two plots, one for the MSE and the other R-squared, to be used in our presentations and reports. As a final step, I combined all our code contributions from team members into a cohesive framework, ensuring consistency and accuracy. When all the results were collected from my teammates, I converted our current plots from Seaborn to Plotly Express and implemented an interactive dashboard using Dash by Plotly. It was both challenging and rewarding, allowing us to present our results in a user-friendly manner. As part of our report, my responsibilities included the literature review, which provided context for our study by delving into existing research. I also wrote the methodology section pertaining to the Random Forest Regressor, handled the data cleaning section, and, of course, this individual contribution report. My contributions span the entire project life cycle, from shaping the research question to presenting our findings. Through Python-based data science techniques, we've shed light on the complex relationship between weather dynamics and the energy sector, contributing to our understanding of energy systems.

8.4 Sam Robbins

Working with my team members was a joy. All tasks were shared very evenly ensuring no-one took on too much work and that everyone would be working hard. Everyone completed all components that was asked of them asking for help from each-other where necessary to ensure a smooth operation throughout the project. Each team member kept each-other up to date with what they were doing meaning that everyone known the entire project in and out.
Progress Report: I aided creation of the project progress report document and I improved the data-sheet for the dataset section within the progress report documentation after receiving feedback.
Progress Report Presentation: I produced the progress presentation introducing the chosen project topic along with initial data analysis with explanation of datasets including access rights. This was aided by the entire team to make sure everyone knew what was happening.
Data Collection: I located a suitable dataset for energy showing consumption of energy by source on a monthly basis as well as finding the associated license and usage.
Data Cleaning: I performed initial data cleaning on the energy dataset allowing comparison with Mo for the most suitable method. I performed cleaning for wind data replacing nan values with the average to allow the models to work with little effect on performance. I helped with the final consolidation of the multiple datasets ensuring the correct time frame was used whilst removing unnecessary data.
Background Research: Researched deeply into the applications of machine learning within electricity and energy consumption sector. I also researched into the applications of different models on similar problems in order to gain a good understanding of the applicability of different models on different problems.
Poster: Created poster layout ensuring formatting was consistent and balanced and wrote several sections. Compiled final versions of the poster.
Final Presentation: Created my associated sections of the final presentation as well as general formatting.
Linear Regression and Comparison to Average: Performed several regression models ranging from linear to polynomial. Multiple linear regression gave the most effective results and linear regression using a support vector regression model was the most reinforced by the literature out of the regression models tested. I also produced an additional predictor that assumed average values for the targets to act as a base for performance comparison between regression models. This gave perspective to the models showing us that the models were giving better results than if they were randomly predicting close the average. A linear support vector regressor was found to be best and so I performed some minor parameter tuning but to little avail. I created a cross validation function that could be applied to all models but this lead to issues as discussed in the method section.
Neural Networks: Learned how to use neural networks. After initially researching into the pytorch library I found using the TensorFlow library more intuitive and used this to apply a general neural network model to the problem. Through more research I decided on the depth of the model, number of neurons in each layer and which optimiser would be optimal. After finding the neural networks were prone to over-fitting, I tuned parameters, but this had little effect. In order to remedy this, I researched into regularisation settling on the dropout regularisation technique. Dropout was applied and results became much more comparable to the other models used in the project.
Report Write-up: For the methodology I wrote the general part applied to all models, linear regression and neural networks methods sections. I wrote the respective results sections for each model and the performance evaluation of metrics results section. I wrote the discussion section with Aman. I proof read the report repaired references and formatting of figures.
Again, The team was a pleasure to work with and it is a shame for the module to come to a close.

Appendix A: Code and Links

XGBoost Lag Feature

```
def add_lag_features(data, num_days_lag):
    Columns = data.columns.values.tolist()
    for col in Columns:
        for i in range(1, num_days_lag+1):
            data[f"{col}_lag_{i}"]= data[col].shift(i)

    return data
```

Link to GitHub Repository :

https://github.com/Sam-Robbins/Weather-and-Energy

Appendix B: Additional Figures

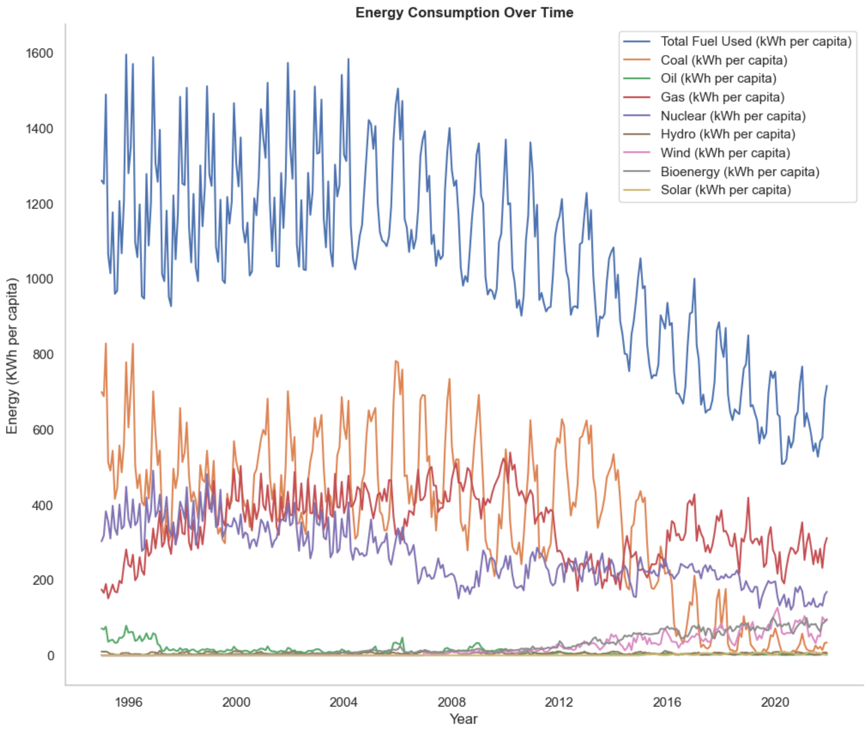
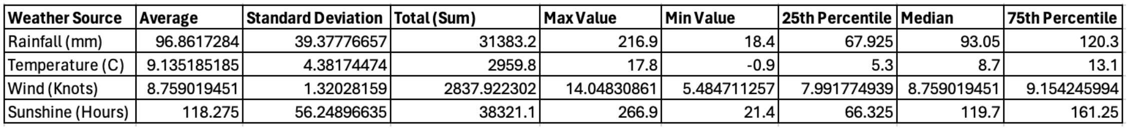


Figure 16: *Electricity sources against time for the full time frame of data used*



Weather Source	Average	Standard Deviation	Total (Sum)	Max Value	Min Value	25th Percentile	Median	75th Percentile
Rainfall (mm)	96.8617284	39.37776657	31383.2	216.9	18.4	67.925	93.05	120.3
Temperature (C)	9.135185185	4.38174474	2959.8	17.8	-0.9	5.3	8.7	13.1
Wind (Knots)	8.759019451	1.32028159	2837.922302	14.04830861	5.484711257	7.991774939	8.759019451	9.154245994
Sunshine (Hours)	118.275	56.24896635	38321.1	266.9	21.4	66.325	119.7	161.25

Figure 17: *Weather Sources summary table*

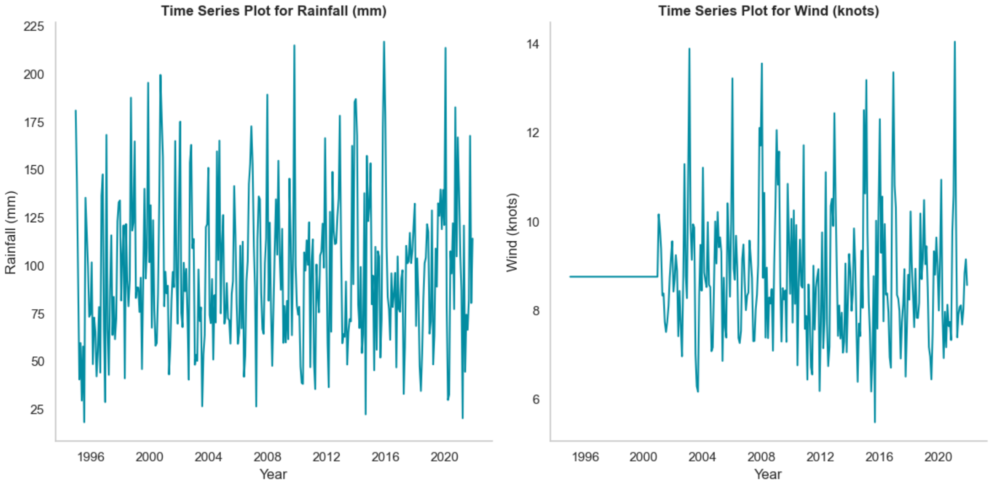


Figure 18: *Rainfall and Wind weather patterns against time for the full time frame of data used*

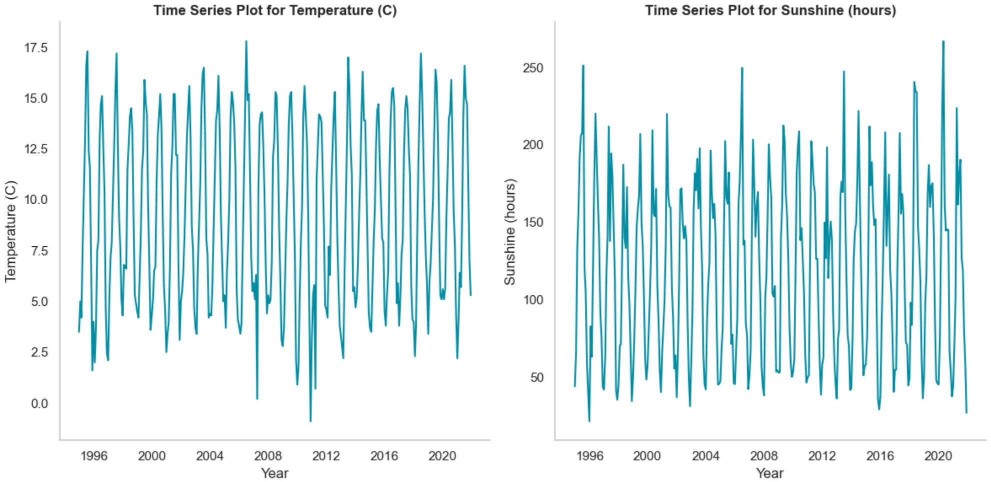


Figure 19: *Temperature and Sunshine weather patterns against time for the full time frame of data used*