|  |
| --- |
|  |
| Using Machine Learning Techniques to Predict Energy Consumption |
|  |

|  |
| --- |
| Usman Kolo Mohammed  4-30-2022 |

Table of Contents

[Introduction 5](#_Toc102176507)

[Background 5](#_Toc102176508)

[Organisation of thesis 6](#_Toc102176509)

[Literature Review 7](#_Toc102176510)

[Introduction 7](#_Toc102176511)

[Regression Models 7](#_Toc102176512)

[Machine Learning Introduction 8](#_Toc102176513)

[Long Short-Term Memory (LSTM) 8](#_Toc102176514)

[Bidirectional-LSTM (BLSTM) 10](#_Toc102176515)

[Gradient Boosting Regression Trees 10](#_Toc102176516)

[XGBoost 10](#_Toc102176517)

[LightGBM 11](#_Toc102176518)

[Facebook Prophet 11](#_Toc102176519)

[Methodology 13](#_Toc102176520)

[Model 13](#_Toc102176521)

[RNN 13](#_Toc102176522)

[LSTM 14](#_Toc102176523)

[Bidirectional LSTM 15](#_Toc102176524)

[XGBoost 16](#_Toc102176525)

[Prophet 17](#_Toc102176526)

[LightGBM 18](#_Toc102176527)

[Loss Functions 18](#_Toc102176528)

[Mean Squared Error 19](#_Toc102176529)

[Mean Absolute Error 19](#_Toc102176530)

[Root Mean Squared Error 19](#_Toc102176531)

[R-squared 20](#_Toc102176532)

[Data Analysis 21](#_Toc102176533)

[Queens Building Dataset 21](#_Toc102176534)

[Preprocess/Cleaning 21](#_Toc102176535)

[Visualize 22](#_Toc102176536)

[Check for stationarity 23](#_Toc102176537)

[Correlation 24](#_Toc102176538)

[Feature Importance 25](#_Toc102176539)

[Hugh Aston Building Dataset 26](#_Toc102176540)

[Preprocess/Cleaning 26](#_Toc102176541)

[Visualize 27](#_Toc102176542)

[Check for stationarity 28](#_Toc102176543)

[Correlation 28](#_Toc102176544)

[Feature Importance 30](#_Toc102176545)

[Gateway House Building Dataset 30](#_Toc102176546)

[Preprocess/Cleaning 30](#_Toc102176547)

[Visualize 31](#_Toc102176548)

[Check for stationarity 32](#_Toc102176549)

[Correlation 32](#_Toc102176550)

[Feature Importance 34](#_Toc102176551)

[Weather Dataset 34](#_Toc102176552)

[Preprocessing 35](#_Toc102176553)

[Visualization 37](#_Toc102176554)

[Feature Selection 39](#_Toc102176555)

[Normalization 39](#_Toc102176556)

[Min-Max Scaler 39](#_Toc102176557)

[Standard Scaler 40](#_Toc102176558)

[Experiments and Results 41](#_Toc102176559)

[Queen’s Building 41](#_Toc102176560)

[LSTM 42](#_Toc102176561)

[BLSTM 42](#_Toc102176562)

[XGBoost 42](#_Toc102176563)

[LightGBM 42](#_Toc102176564)

[Prophet 42](#_Toc102176565)

[Hugh Aston 41](#_Toc102176566)

[LSTM 42](#_Toc102176567)

[BLSTM 42](#_Toc102176568)

[XGBoost 42](#_Toc102176569)

[LightGBM 42](#_Toc102176570)

[Prophet 42](#_Toc102176571)

[Gateway House 41](#_Toc102176572)

[LSTM 41](#_Toc102176573)

[BLSTM 41](#_Toc102176574)

[XGBoost 41](#_Toc102176575)

[LightGBM 41](#_Toc102176576)

[Prophet 41](#_Toc102176577)

[Conclusion 42](#_Toc102176578)

[References 43](#_Toc102176579)

# Introduction

## Background

Buildings are the largest energy consumer in the United Kingdom, accounting for 40% of total energy consumption [1]. The consumption of electricity in Higher Education (HE) facilities accounts for most of the sector's carbon pollution [2]. Once the Higher Education Funding Council of England (HEFCE) mandated a 43 percent reduction in carbon emissions by 2020 for all participating institutions compared to the baseline year 2005, energy management solutions have quickly acquired traction in the UK HE sectors [3]. Most of England's member institutions now have a dedicated energy management team dedicated to achieving this goal. Figure 1 depicts the change in Dioxide (CO2) emissions in the sector from 2008/09 through 2014/2015. It shows that electricity usage is the largest contributor, accounting for 63%, followed by natural gas, which accounts for 33.3% [1,4,5].

Building type, building age, occupancy, working hours, kind of equipment fitted, and weather patterns are all variables that impact energy usage in university buildings. Academic buildings (42%) and administrative buildings (26%) take up about 68 percent of the space on a typical university campus in England [6]. In 2014/15, non-residential facilities at English universities utilised around 80% of overall energy consumption, while residential buildings consumed 20% of total energy consumption [4].

With shares of 63 percent and 33 percent, respective, the previously stated data shows clearly how gas and electricity use are the two biggest carbon-emitting sources in England's HE sectors [4]. The modest decline in carbon emissions in the industry implies that universities ought to investigate more energy-saving alternatives as well as improve their energy management systems.

One of the essential parts of a successful energy management system is monitoring and analysing building energy consumption trends, which aids in analysing the facility's operational behaviour under various conditions. It also aids in the detection of unintended energy waste under specified conditions. If energy consumption data is properly extracted and stored, the relationship between the data and various variables such as temperature, humidity, number of occupants, and so on may be explored, and future energy projections can be formed using all these variables. Another significant advantage is that forecasted energy consumption data might be utilised to forecast realistic future energy budgets. Universities' energy management teams oversee monitoring, evaluating, and keeping data on energy consumption in their buildings. They oversee setting up realistic energy consumption predictions and identifying options for energy conservation to build their energy budget predictions for the years ahead.

The ability to accurately estimate energy use is important to the effective application of energy management systems. Due to a major lack of forecasting and the adoption of less efficient forecasting methodologies for planning, many businesses have struggled to control energy use and budgets [7,8]. Forecasting aids in assessing present and future economic conditions to guide the organization's policies and decisions. It's a technique that uses past and current data to forecast future information. A credible prediction system can assist financial and energy management teams at universities in establishing objectives and strategic goals and can even be used as part of its annual budgetary process [9].

Researchers have extensively employed diverse predictive techniques including MR, ANN, and GA for energy consumption prediction for various building types in various areas [10]. MR is an easy, reliable, and rapid technique among such [11,12]. Several researchers have employed the MR approach in their studies [11,12,15], but all such MR models anticipate the energy consumption of a single building or an area and require a large amount of input data. Energy managers and their staff are often busy, thus a single, trustworthy, and rapid model for diverse building categories would be preferable to multiple forecasting methods.

This study intends to assist the university's energy management teams by creating a fast and easy predictive model that employs a variety of methodologies. Every manager should be able to anticipate the hourly, daily, and monthly energy usage of the facilities using these methodologies. Three buildings on DE Montfort University's main campus were chosen for this purpose, and their past energy usage data was used to generate the forecasting models.

## Organisation of thesis

There are five chapters in this thesis, as well as an opening abstract. The abstract will give the reader a quick overview of the work that went into creating this dissertation. The foundation of the dissertation is discussed in depth in the first chapter. The chapter also discusses the energy industry's present issues, the purpose for energy prediction, and a brief explanation of the IAC programme. The second chapter looks at various methods for predicting energy usage in buildings. After a brief overview, the chapter discusses how each approach (regression, support vector machines, random forests, and artificial neural network (ann) performed in estimating energy. The methodologies employed in this thesis' case study from the IAC database are detailed in the third chapter. The workings of regression, support vector regression and random forests in parameter prediction are all detailed in this chapter. The fourth chapter explains each variable in the dataset utilised for prediction. The chapter also includes the many equations used in the case study and explains how to apply the methodology from chapter 3 to the dataset utilised in this thesis. The fifth chapter examines the model's performance and describes the results obtained from the model constructed in chapter 4. The chapter also examines a few conclusions drawn from the data. The dissertation is concluded in the last chapter, which also specifies prospective future research.

# Literature Review

## Introduction

The goal of this chapter is to provide a thorough examination of the efforts undertaken in the field of energy prediction. This chapter will cover a variety of subjects all connected to energy forecasting models and the results that determine which method is the best. In the past, best-fit regression equations were utilised for prediction and analysis [14]. This can be thought of as a "go-to" strategy for constructing equations from historical data and predicting future key variables. To better forecast the future, modern machine learning techniques are replacing or collaborating with traditional methodologies such as regression. This chapter will discuss previous work on regression, and machine learning techniques such as decision trees, Facebook Prophet, and RNNs specifically for energy prediction.

## Regression Models

[10, 15] described how regression can be used to predict energy use. According to [10], the amount of energy consumed by buildings supports the need for thorough research and modelling in this field. According to the article, quantitative methods are a particularly convenient option for constructing energy models when its user only has access to historical data but not the multiple values necessary for technical equations. When compared to other quantitative methodology, linear regression was determined to be a comparatively simple and accurate application. Hourly and daily records from a residence dedicated to research were used in this study. Because of the time difference, the researchers were able to investigate the impact of data collection frequency on the model's precision. External temperature and sun radiation were significant variables. The dependent variable was energy consumption. Simple and multiple linear regression, as well as quadratic linear regression, were all examined. It was carried out to check if the quadratic regression's extra depth was supported by a higher quality of performance.

The time interval was discovered to be a critical aspect in determining the model's quality. The model's performance improves with the length of time. To clarify this, the researchers state that when data is obtained over a shorter period, energy usage anomalies reveal large disparities. When a larger amount of data is acquired over a longer period, the errors average out over time. Using a multiple linear regression model, the coefficient of determination was enhanced. However, the RMSE suffered because of this method. When both criteria were considered, it was discovered that multiple linear regression provided the best overall quality of energy prediction. Daily time intervals also yielded the most accurate model parameters.

A multiple regression model for predicting heating energy demand was presented in [15]. Heating energy demand, according to the authors, is a crucial estimate. It is used during the design phase of a building to anticipate how much energy will be required for space conditioning over its lifetime. The global energy loss coefficient of a building, south equivalent surface, and temperature difference were used as independent variables. [15] is a simpler model than [10] because only three variables and a single prediction model were utilised. [15] also adds that in the event of big datasets, such as the one used in this work, a regression can be employed with reasonable success, even though the fact that it is simpler and easier to construct. The model in this article was developed using the "black box" idea. When input and output variables are known, and the user is required to fit in the best curve possible (also known as "black-box" owing to the unknown nature) to build a generalised relationship between dependent and independent factors, this approach is applied. Though least squares estimation is the most popular method, it can occasionally result in errors that are not normally distributed, which is a strong validation of any curve fitting model [15]. As a result, we employed an incrementally reweighted least-squares method. To lessen the impact of a residual anomaly, this approach modifies the weight of the coefficients in the regression model. Ultimately, this method of determining the best fit curve yields a better least square estimate.

The model was evaluated on 17 blocks of apartments once it had been trained. With an  of 0.9744, the model was determined to be quite efficient. Also, it was discovered that 90% of the calculated results exhibited relative errors of 20% or less. [15] also analyses the model against several dynamic solutions, claiming that such a proposed approach is quicker and produces similar outcomes. Regression is the optimum strategy in terms of model quality and efficiency when constructing models for datasets with a lesser set of variables, according to [15].

## Machine Learning Introduction

Machine learning is described as the application of computational methods that have been trained on historical data to assist in the decision-making process for a specific system. These tools are typically used to increase performance or create precise forecasts. Machine learning approaches are currently being utilised in a variety of industries to make predictions using historical data, either as a substitute or in conjunction with traditional regression/statistical models [16]. [16] is a good place to start reading about machine learning's real-world applications. Several machine-learning techniques are now being utilised to estimate energy usage, including but not restricted to RF [17], SVMs [18, 19], and NNs [20, 21, 22]. Several studies have found that ML algorithms are at least as effective as traditional methods. This chapter delves into such publications in-depth.

## Long Short-Term Memory (LSTM)

Given the availability of studies that use alternative networks, the far more effective methods for forecasting electricity usage are LSTM networks. As a result, the research presented in [27] investigated the usage of multiple LSTM designs for projecting electricity use in the near to medium term. The ideal number of hidden layers with time lags was determined using a GA. The approach's appropriateness was tested using data from France's usage. Bedi et al. [26] provided a methodology for analysing past data's long-term connections as well as short-term trends in segmented data. Later, utilizing electric load data from India, LSTM was implemented with a shifting frame. DRNN, ANN, and SVR were all surpassed by the model constructed. [28] presents the analysis of power prediction using temperature as an independent factor. They combined power demand data with some other factors such as temperature, air pressure, and humidity to predict power demand usage data.  For the short to medium range  (24 and 48 hours, 7 and 30 days), the prediction is used. Performance measurements like RSME and MAPE were used to compare the proposed model to other conventional methods. The analysis indicated that LSTM surpasses competing algorithms when it comes to boosting predictive performance. Using a MATLAB toolkit, the LSTM system was autonomously tuned. For many prediction horizons, the results are then compared to that of ARMA, SARIMA, and ARMAX.

External variables were also given to the LSTM by Kwon et al. [29]. These hyperparameters were configured using a heuristic methodology. The electricity system administrator in Korea utilised 2 years to analyse the model, with only an error of 1.5%. The LSTM presented by [30], on the other hand, provided a data dimension reduction to reduce computing costs. To evaluate the approach's effectiveness, the authors devised two groups of trials. The suggested technique outperformed ANN, ARMA, and ARFIMA in comparisons. The suggested LSTM technique was shown to have an RMSE of 19.7% lesser than that of the benchmark feed-forward neural network. Subsequently, [31] suggested the CVOA, which has been utilised to improve the hyper-parameters of an LSTM network. The findings provided surpassed a large variety of deep learning techniques combined with well-known optimization strategies. As a baseline, data on Spanish power use has been employed. [32] proposed a multi-layer bi-directional RNN centred on LSTM and GRU to estimate electricity usage. The researchers beat the findings of ANN and SVR when they looked at peak usage and seasonality individually. Pegalajar et al. introduced in recent years three variants of RNN to anticipate Spanish energy demands, comparing the results against a wide range of ML models and finding that they outperformed every one of them [33].

Utilizing the LSTM algorithm and raw data from a housing complex, Kong et al. [34] established a framework for estimating short-term load. For most datasets, the suggested LSTM architecture had the best prediction performance. Based on the LSTM method, Jiao et al. [35] created a model for forecasting the demand for non-residential buildings. The suggested framework was built using real data from 48 non-residential buildings, which included energy consumption statistics. Kim and Cho [36] used a mixture of LSTM and CNNs to design a model for precisely estimating the power usage of a housing complex for a steady electrical supply. CNNs were designed to obtain the features of variables that affect power use in the research, which were initially designed for image analysis. LSTM may detect details in time-series data that have an abnormal pattern. Through this work, a CNN–LSTM model was created by merging the properties of the two algorithms, and electricity consumption was forecasted almost accurately, which was formerly hard to forecast. Even with minute, hourly, daily, and weekly data, a high level of predictive performance was discovered.

Khafaf et al. [37] suggested an LSTM neural network model for forecasting energy consumers' 3-day energy use clusters. Wang et al. [38] used the LSTM neural network to estimate power usage and identify grid irregularities. Regarding energy consumption forecasting in residential and commercial buildings, Khan et al. [39] used a hybrid CNN with an LSTM autoencoder. Regarding everyday natural gas consumption forecast, Wei et al. [40] developed a hybrid single spectrum analysis and an LSTM model. Singaravel et al. [41] examined the effectiveness of LSTM in the planning phase of a building. With four alternative versions of the LSTM model, 201 design scenarios were analysed. LSTM models are found to have superior accuracy and processing time than ANN models. To estimate the power usage of air-conditioning systems, Zhou et al. [42] presented an LSTM model. The number of iterations, the size of the time series feeding the LSTM, and the learning rate were all tested repeatedly to identify the ideal hyper-parameters.

Furthermore, to increase LSTM's predictive performance, hybrid evolutionary optimisation approaches such as GA and PSO are merged with it. To find the optimum weight matrix or component of the LSTM hyper-parameters, the majority of GA and PSO are used. He et al. [43] presented a hybrid short-load forecasting system combining variational mode decomposition and LSTM networks, with the LSTM network's hyper-parameters tuned by using the Bayesian Optimisation algorithm. For home energy usage forecasting, Kim et al. [44] developed an LSTM network. To determine the best hyperparameters, including learning rate, layer size, and dropout rate, the PSO is used. Yang et al. [45] suggested a hybrid prediction model that combined extreme learning, RNNs, and SVMs. To choose the best weight matrix between these networks, PSO is used. Guo et al. [46] presented an LSTM neural network-based short-term prediction method that captures real-time responses into account. To get the appropriate weight matrix for LSTM, the enhanced GA is applied. The LSTM network's architecture, on the other hand, is set based upon that developer's expertise. Regarding hourly natural gas demand forecasting, Su et al. [47] developed a hybrid wavelet transform and an LSTM model. Trial-and-error is used to determine the amount of LSTM layers, whereas GA is used to optimise the number of neurons for each LSTM layer.

## Bidirectional-LSTM (BLSTM)

Deep neural networks based on BLSTM have been extensively used in speech recognition and text classification, but they are seldom used in time series prediction or stock market forecasting. RNNs aid in the detection of relationships in sequential data. Bi-LSTMs are a type of RNN that can model data forward as well as backwards. This incorporates several previous stock prices as well as potential future stock values. M. Jia et al. [25] established a methodology for predicting the future price of a stock using a BLSTM neural network. The writers used the GREE stock's statistical data. From January 2017 to May 2019, they compiled data for 568 days. There were 14 different features in the data, which included open, high, close, volume, and so on. The data was pre-processed and standardised. For the forecast, the close price was used as a baseline. A one-way and two-way LSTM were used to process the pre-processed data. Overfitting can be avoided by using a technique called dropout. The neural network was subjected to the dropout approach. The suggested BLSTM was assessed to both the ARIMA and a standard LSTM model. The RMSE, MAE, and deviation result of the proposed system were determined to assess its accuracy. The suggested approach surpassed both the ARIMA and the LSTM models, as evidenced by the findings. The RMSE and MAE also were lowered by 24.2% and 19.4%, correspondingly.

[23], on the other hand, adopted BLSTM for energy load prediction. The performance of various architectures was examined, including BLSTM, multilayer LSTM, and decoder-encoder architecture. The BLSTM structure was found to have the best results. [24] also used BLSTM for network-wide traffic speed forecast. Both linear SLSTM and BLSTM neural networks were examined in this study. It investigated several LSTM structures and found that two layers of stacked BLSTM surpass other LSTM-based designs. Employing multiple tests to assess BLSTM with SLSTM, we study and assess the integration of a suggested methodology using BLSTM into financial time series forecasting.

## Gradient Boosting Regression Trees

GBRT variants have been pitted against a slew of newer models. Random Forest [56] , XGBoost [56][57], CatBoost [58], AdaBoost [59], SVM [60], MLP [61], and CNN and GRU framework [62] have all outperformed GBRT. [63] advocated evaluating 7 algorithms per day during the forecast, including GBRT and XGBoost, and then using the most effective for the next day's forecast. One suggested incorporation of the model is to use it as a foundation for a stacking model that includes RF, XGBoost, SVR, and kNN to increase performance and generalisation [65]. Investigating feature selection approaches such as Mutual Information, F-regression, Elastic Net, and Recursive Feature Elimination in tandem with the GBRT model [64] is another field of research that is beyond the scope of this paper.

### XGBoost

[65] developed a new method for anticipating electricity demand. They took daily power load data and translated it to weekly power demand data. This expands the number of features that can be used to forecast load for a lag parameter. For selecting features from the transformed data, the XGBoost method was utilised. After that, the model was trained, yielding a minimum MAPE of 10%, 97% accuracy, and an MAE of 88.90%.

When used in the literature, the XGBoost approach has had a lot of success.   XGBoost or XGBoost hybrid systems surpassed other methods [57], [66], [67], [68]. With its capacity to execute parallel processing, XGBoost is incredibly effective [67]. Before training, XGBoost organises the data, stores it in a block structure, and then uses it in later phases. This increases performance by drastically decreasing processing, as grouping the values of the attributes are amongst the most time-consuming phases in decision tree learning. As a result, in the Wang et al. [56] analysis, XGBoost had the optimum performance. Another aspect of XGBoost is its built-in capacity to choose and save important features; as a result, additional feature selection algorithms aren't always necessary by the XGBoost algorithm and can even degrade XGBoost efficiency [70].

XGBoost was used in conjunction with other techniques by several academics. CEEDMAN would be used initially to de-noise the original data by dividing it into 12 samples, according to Lu et al. XGBoost surpassed SVM with PSO, Least-squares SVM, CEEMDAN Random Forest, and Radial basis function NN in the 12 samples, and the forecast results were then averaged and denormalized. [73]. Pairing EMD and ARIMA with XGBoost [74], along with solely using ARIMA [71], and merging XGBoost with k-means on Similar Days selection [69] are two further instances. All the XGBoost-based models in the prior articles surpassed other algorithms in contrast. Different architectures such as the CNN and GRU frameworks [62], as well as an ANN [72], were able to best XGBoost in a few articles.

### LightGBM

A concurrent IDS built on a PID method and LightGBM was proposed by Jin et al. [78]. To accomplish response time outside of reducing attack detection performance, the suggested framework employs two main methods. To begin, the intrusion sensor is a LightGBM. Next, traffic data is properly analysed using PID. Rapid IDS is built on PID methods that have expenses in terms of collaboration and communication. Furthermore, with a connection rate of up to 1.26 Gbit/s, the suggested framework is stable.

When dealing with non-linear traffic data, machine learning is more precise than mathematical statistics. Nonetheless, the lack of understandability of prediction outputs limits ML. A variety of boosting techniques inspired by the GBDT have recently been discovered. XGBoost [75] and LightGBM [76] are two examples. Due to its parallel learning, robustness to anomalous values, and adaptability to missing values, they have been used and coupled with other models in a variety of domains. To predict ultra-short-term wind energy, Ju et al. [77] suggested a hybrid model based on CNN and LightGBM, and the findings confirmed that the inclusion of LightGBM surpassed effectiveness and precision.

## Facebook Prophet

[48] present Prophet as a young technique for predicting that has a lot of promise for use in power demand prediction. In recent years, various applications of this concept have been discovered. In the case of Bitcoin prediction, Yenidogan et al. [49] examined two methods: ARIMA and Prophet. The results suggest the Prophet model had a precision of around 94.5%, making it significantly higher than ARIMA's 68%. Furthermore, Ashwini Chaudhari [50] forecasted the costs of currencies like Bitcoin, Litecoin, and Ethereum using three models: ARIMA, Prophet, and LSTM networks. The findings show that using LSTM and Prophet yielded very accurate results for the 3 cryptocurrencies, ranging from 93% to 99% while using the ARIMA model yielded just 82% to 66% accuracy. Furthermore, Bianchi et al. [51] used raw data from an Italian utility business to compare thermal short-term demand forecasting methodologies utilising the ARM, NARM, and Prophet. In terms of short-term predicting, the ARM outperformed the other models. Das [52] conducted a comprehensive study in which five alternative forecasting models (SES, Dynamic Harmonic Regression, NN,  ARIMA, and Prophet) were utilised for wind speed prediction in two Indian states (Tamil Nadu and Maharashtra). The greatest results came from the neural network. The Prophet framework, on the other hand, produced encouraging results and was suggested for future purposes.

The evaluation of random forests and Facebook's Prophet in predicting daily flow up to 7 days in advance in a river in the United States is examined by Papacharalampous & Tyralis [53]. These prediction systems employ historical streamflow data, with random forests also using historical rainfall data. They employ a naive approach based on previous streamflow observations and an MLR model that incorporates identical data as random forests. The findings demonstrate that random forests surpass the naive technique overall, whereas Prophet surpasses it over prediction timeframes greater than 3 days.

Earlier studies [54] employed it to forecast sales, and the findings were presented using the MAPE level for sales forecasting of various categories of items. On a quarterly forecast, they were able to attain a MAPE of little less than 30% for 70% of the goods. For this study, standard seasonality trends were applied. The WMS manages a complicated series of processes known as warehousing. The idea of smart WMS is outlined in [54]. The implementation of Facebook's Prophet algorithm for sales forecasting was outlined by Zunic et al. [55] as part of the smart WMS notion and improvement of supply companies. In several of Bosnia and Herzegovina's major facilities, the notion of smart WMS and sales forecast has been proven in actual scenarios and with real data. According to the preceding published papers, no study has yet been done on the Prophet model's accuracy in long-term energy load prediction.

# Methodology

## Model

First, we'll go through the RNN and its LSTM structure. Following that, the LSTM will be completely discussed, including its bi-directional properties. Then we'll go into XGBoost and its capabilities in tackling this problem in greater depth. The Facebook Prophet algorithm will be explored, followed by an examination of the LightGBM.

## RNN

Recurrent Neural Networks are robust and effective classes of neural networks ranking among the top algorithms due to them being one with internal memory.

They are an old deep learning algorithm like many others developed in the 80s, but the true potential was discovered only recently. RNNs have been propelled to the forefront through the increased processing power and a large amount of data available today also with the advent of LSTM in the 90s.

Because of the internal memory, they retain vital information about the inputs received, making it very precise in predictions. Therefore, it is the trusted algorithm for time series data, speech, text, weather, video, and much more sequential data. Unlike, other algorithms RNNs have a more comprehensive grasp of the sequence and its context.

How do they work? The data is passed via a loop. After a choice has been made, it accounts for both the current and past inputs. This is shown in the figure below.

Diagram

Description automatically generated

Unlike a feed-forward neural network, RNNs add the past to the present. This is explained when we feed to the word “apple” as an input; the feed-forward would forget the previous information as it progresses, making it impossible to predict what letter comes next. RNNs remember information by making duplicates of the output and feeding it to the network.

The recurrence is a feedback link that has been created to allow past information to be sent back and received again. Figure 2 illustrates this point. RNNs can be layered and enlarged to better issue modelling while maintaining the previously improved time-series forecasts. Figure 3 depicts a more expanded form of RNN.

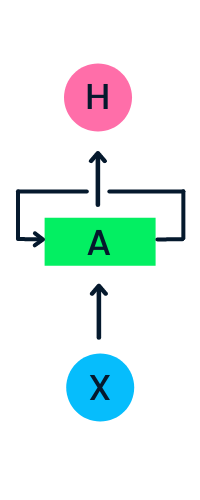


Figure 2: Simple RNN structure with a hidden layer

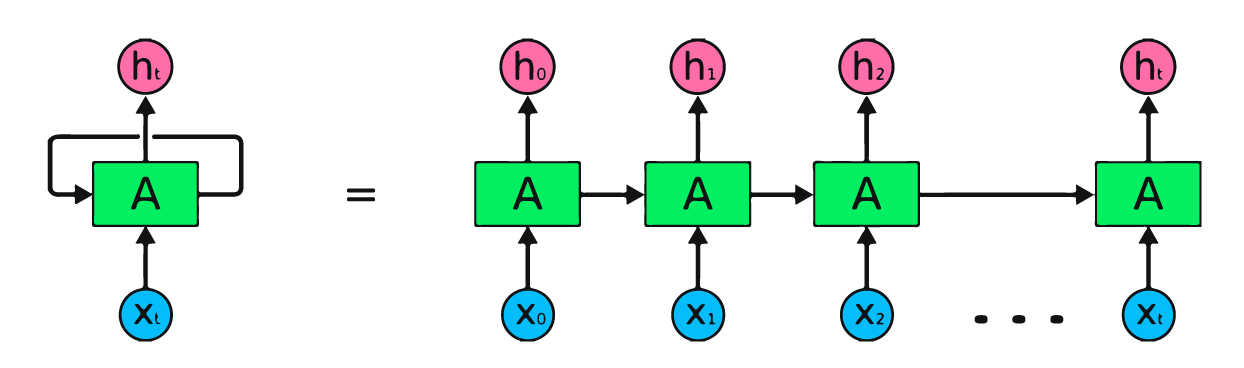


Figure 3: RNN with one hidden layer expanded over time.

### LSTM

Long short-term memory networks (LSTM) are a type of Deep Learning network. It is a class of RNN that can learn long-term connections, which is useful for solving sequence prediction issues. Apart from single data points like photos, LSTM has backpropagation, which means it can analyse the complete sequence of data. This is useful in natural language processing, machine learning, and other areas. The LSTM is a type of RNN that performs exceptionally well on a wide range of issues.

An LSTM model's central role is played by a memory cell called a 'cell state,' which maintains its state across time. The horizontal line which passes through the top of the diagram below represents the cell state. It can be compared to a conveyor system on which data just passes, unmodified.

In an LSTM, information can be provided to or withdrawn from the cell state, which is controlled by gates. These gates allow information to move in and out of the cell if desired. The method is aided by a pointwise multiplication operation as well as a sigmoid neural - network layer.

The sigmoid layer outputs integers between 0 and 1, with 0 indicating that "nothing should be let through" and 1 indicating that "everything should be let through". This is represented in Figure 4.

Diagram

Description automatically generated

Figure 4: This image visualizes a single LSTM neuron.

LSTM can address a variety of issues that prior learning algorithms, such as RNNs, couldn't. Long-term temporal connections can be efficiently handled by LSTM without the need for a lot of optimization work. This is used to deal with high-end issues.

### Bidirectional LSTM

Bidirectional-LSTM allows any neural network to store sequence information both in forwarding and backward directions.

The data runs in two channels in a bi-lstm, which distinguishes it from a conventional LSTM. We can have input move in one way, either forwards or backwards, with a normal LSTM. However, with bi-directional input, we can have the flow of information in both channels, preserving both futures as well as the past. Let's consider the sentence for a deeper understanding.

The vacant area in the line "guys visit..." cannot be filled. However, if we have an upcoming sentence like "the guys came out of the pub," we can readily anticipate that previous vacant space and then have the model do the same thing, and bi-LSTM enables the network to do so.

Diagram

Description automatically generated

The information flow from forward and backward levels can be seen in the figure. When jobs requiring sequence to sequence are required, BI-LSTM is typically used. Text categorization, natural language processing, and predictive models can all benefit from this type of network.

## XGBoost

This is a gradient boosting application that utilizes a decision-tree-based group ML model. Artificial NNs surpass most other algorithms or methodologies in prediction issues involving complex data (pictures, text, and so on). Decision tree-based algorithms, on the other hand, are now rated best-in-class for small-to-medium structured/tabular data.

Diagram

Description automatically generated

Evolution of XGBoost from Decision Trees

The XGBoost algorithm was created as part of a University of Washington research study. In 2016, Tianqi Chen and Carlos Guestrin submitted their article at the SIGKDD Conference, which ignited the ML industry. Since its inception, this algorithm has been acknowledged for not only topping a slew of Kaggle contests but also serving as the brains behind several cutting-edge industrial applications. As an outcome, the open-source projects have a robust ecosystem of data scientists collaborating with them, with 350 members and 3,600 contributions on GitHub. It sets itself apart through this means:

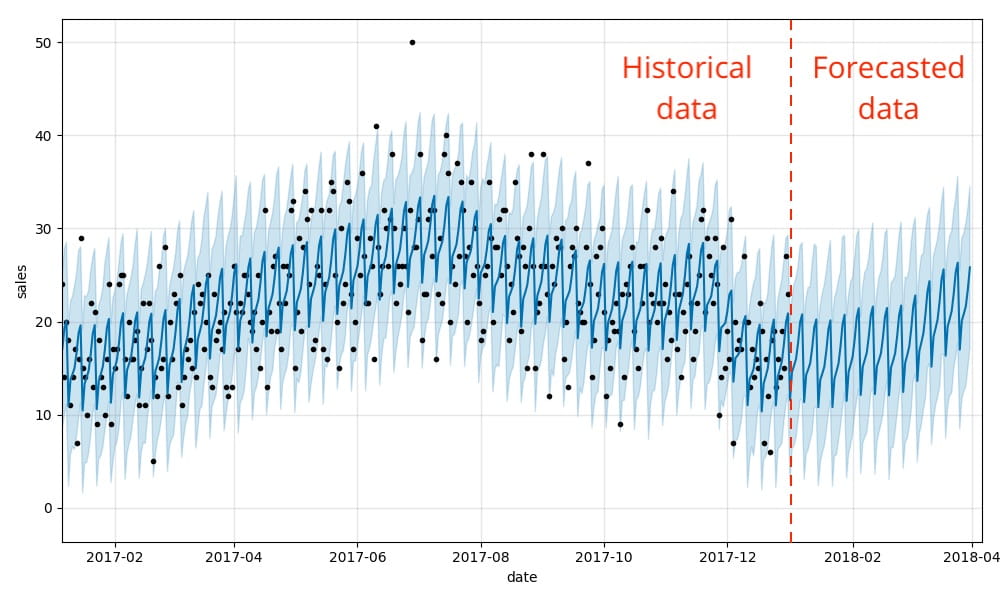
* Regression, classification, ranking, and user-defined prediction issues can all be solved using this tool.
* Very portable i.e., runs on popular Operating Systems (Windows, Linux, etc).
* Supports all major programming languages
* Supports cloud integration i.e., Azure, AWS, and other ecosystems.

With these facts, it does not mean the XGBoost algorithm should be used for all ML problems. A good rule of thumb is to explore both new and old frameworks to tackle the desired problems. It will only be a matter of time before a new model framework emerges that outperforms XGBoost in respect of predictive accuracy, adaptability, explainability, and practicality. XGBoost, on the other hand, will continue to rule the Machine Learning field till a serious opponent emerges.

## Prophet

Prophet is a time-series predicting procedure based on an additive model that fits non-linear trends with annual, weekly, and daily seasonality and holiday impacts. It performs effectively with time series with substantial seasonal influences and past data from multiple seasons. Prophet is forgiving of missing information and trend changes, and it usually manages anomalies efficiently.

Facebook's Core Data Science team published Prophet as open-source software. It may be downloaded from [CRAN](https://cran.r-project.org/web/packages/prophet/index.html) and [PyPI](https://pypi.org/project/prophet/).



Example of the Prophet algorithm output

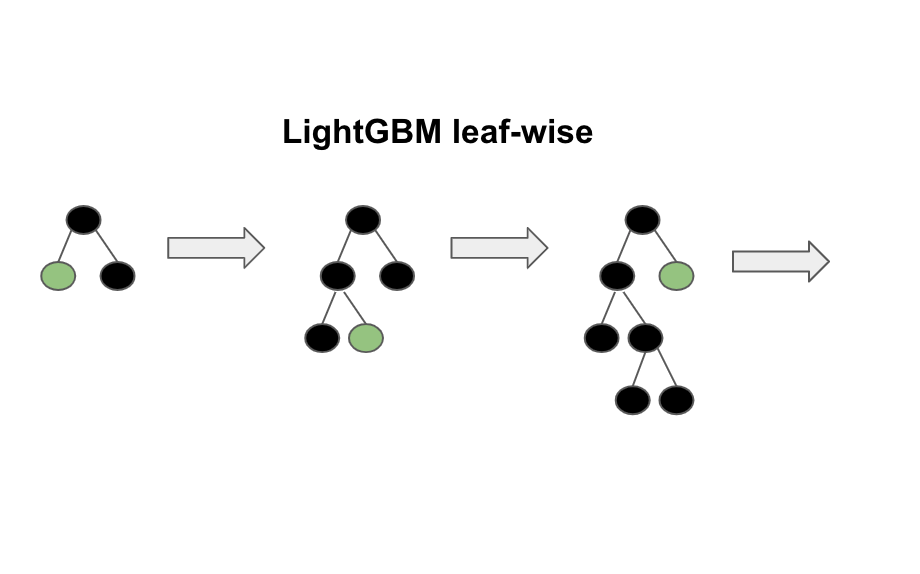
Prophet is utilized in a variety of Facebook apps to generate accurate predictions for strategy and goal setting. In most circumstances, we've found it to outperform any other strategy. We use Stan to fit models so that you may get forecasts in a matter of seconds. With no effort, you can get a fair prediction on sloppy data. Anomalies, missing information, and significant improvements in your time-series data are not a problem for Prophet. The Prophet method gives customers a lot of options for tweaking and adjusting predictions. By combining your subject expertise with human-interpretable variables, you can enhance your predictions. The Prophet process has been coded in R and Python, however, the fundamental Stan code for fitting is the same across both. To get predictions, use any language you're most confident with.

## LightGBM

This is a gradient boosting concept that includes tree-based learning methods, which are regarded to be quite a strong processing algorithm. This is thought to be a quick-processing algorithm.

Whilst trees of other algorithms develop horizontally, the LightGBM algorithm develops vertically, which means it expands leaf-wise while many other algorithms develop level-wise. To expand, LightGBM selects the leaf with the greatest loss. When expanding the same leaf, this can reduce loss more than a level-wise method.

Existing algorithms are finding it increasingly difficult to provide quick results as the number of data grows exponentially. For its computational power and ability to deliver findings quickly, LightGBM is dubbed "Light." It uses less bandwidth to operate and can handle massive volumes of data. Since the goal of the method is to achieve high accuracy while also bracing GPU leaning, this makes it the most often utilized technique in Hackathons.



Diagrammatic representation of Leaf-Wise Tree Growth

Unlike many other boosting techniques that develop trees level-by-level, LightGBM breaks the tree leaf-by-leaf. It grows the leaf with the greatest delta loss. The leaf-wise approach has a smaller loss than that of the level-wise algorithm because the leaf is stationary.

LightGBM isn't suited for small samples. Because of its sensitivity, it can quickly overfit little data. It's suitable for data with more than 10,000 entries. There are no predetermined criteria for considering whether to use LightGBM. This can be utilized for enormous amounts of data, particularly when great accuracy is required.

## Loss Functions

MAE, MSE, RMSE, and R2 are some of the evaluation metrics employed. The MAE indicates how well the model performs on most data by describing the mean error of the model's prediction. Because RMSE is derived from MSE, which gives significant weights to outliers, it is a good measure for predicting outliers. R2 is a metric that describes how effectively a model can represent and fit data. Naturally, the MAE, MSE, and RMSE should be as near to zero as feasible, and R2 should be as close to one as possible.

### Mean Squared Error

The degree of inaccuracy in statistical models is measured by the mean squared error (MSE). The average squared difference between observed and expected values is calculated. The MSE equals zero when a model has no errors. Its value rises as the model inaccuracy rises. The mean squared deviation (MSD) is another name for the mean squared error.

Where:

= number of data points

= observed values

= predicted values

### Mean Absolute Error

Mean Absolute Error (MAE) is the most basic measure of forecast accuracy. It is just the average of the absolute mistakes, as the name implies. The absolute error is defined as the absolute value of the difference between the predicted and actual value. It is a metric that quantifies the precision of continuous variables. This shows us how large of an error we may expect on average from the forecast. It is a statistic that assesses the average magnitude of errors in a set of predictions without taking their direction into account. The Mean Absolute Error is the average of the absolute differences between prediction and actual observation over the test sample, assuming that all individual deviations are equally weighted. The average model prediction error is expressed in units of the variable of interest in both Mean Absolute Error and Root Mean Square Error. The Mean Absolute Error and the Root Mean Square Error both have a range of 0 to 1 and are unaffected by error direction.

Where:

= prediction

= true value

= total number of data points

### Root Mean Squared Error

The root mean square error (RMSE) is the residuals' standard deviation (prediction errors). Residuals are a metric of just how far measured values are from the regression line; RMSE is a measure of how to spread out these residuals. In other words, it indicates how tightly the data is clustered around the best fit line. In climatology, forecasting, and regression analysis, root mean square error is widely used to check experimental results. When normalized observations and forecasts are utilized as RMSE inputs, the correlation coefficient has a direct link. If the correlation coefficient is 1, for example, the RMSE is 0 since all the points are on the regression line (and therefore there are no errors).

Where:

= variable i

= number of non-missing data points

= actual observations

= predicted observations

### R-squared

The coefficient of determination, or , is a statistic that indicates how well a model fits the data. It is a statistical tool about how well the regression line resembles the actual findings in the context of regression. When a statistical approach is used to forecast future outcomes or test hypotheses, it is critical to remember this. The overall sum of squares is the total of the distance the data is away from the mean all squared, and the sum squared regression is the sum of the residuals squared. It will take the value of 0 or 1 because it is a percentage.

# Data Analysis

There are three data sets, each for every building, namely the electricity data set, water data set, and the gas data set collected from the University energy manager. Each dataset contains 48 columns of half-hourly readings being recorded and stored for later analysis. This chapter would discuss the analysis and preprocessing carried out before model training can be carried out. 4 steps are necessary to complete the analysis process are the cleaning/preprocess phase where the data is reformatted and any null values are removed, the decomposition phase where we take a closer look at the data and denote any trends occurring, checking if the data is stationary (without trends) or not which is expected in the case of this project, and finally prepare the dataset for training. The next subsections will go through analysing the building datasets and noting any findings.

## Queens Building Dataset

### Preprocess/Cleaning

A black screen with white text

Description automatically generated with low confidence

Queen’s Building Electricity Reading before cleaning

As the figure above shows, the raw data collected is recorded every 30 minutes from the assigned meter and is not comprehensive to understand, so a preprocessing is essential.

A picture containing table

Description automatically generated

Queen’s Building Electricity Reading after cleaning

After cleaning the data, it is now easy to work with and we can move to the next step in our analysis.

### Visualize

Visualization of the data can be achieved in various ways but to save time we will call upon the statsmodel library which is used in carrying out statistical tests. To use the library, we must resample or data into a daily format giving us the figure.

A picture containing graphical user interface

Description automatically generated

Electricity Daily Dataset

From the figure above we can see that the library was able to display 4 different plots, the original data, the trend of the data, the seasonality, and the residuals (leftovers). The seasonality fails to get detected by the library; a further resampling is essential to observe the seasonal trend.

Graphical user interface

Description automatically generated

Electricity Monthly Dataset

From this new figure we can see a more defined seasonal trend and it is exhibiting a clear downward trend as the year increases. The seasonality also shows more usage in the winter and spring months and significantly less usage in the summer months.

### Check for stationarity

Using the above figure, we can see that our data is non-stationary which increases the desired model’s predictive accuracy. Another way of testing is by performing an Augmented Dickey-Fuller (ADF) test, a unit root test, which is a sort of statistical test. This test is conducted under certain conditions:

* Null Hypothesis: Data is non-stationary
* Alternate Hypothesis: Data is stationary

This test may show that the series is non-stationary if the null hypothesis fails to be rejected. The condition to reject the null hypothesis is:

* If the test statistic < critical value and p-value < 0.05

|  |  |  |  |
| --- | --- | --- | --- |
|  | Test Statistic | P-Value | Critical Values |
| Electricity | -2.983 | 0.036552 | 1% -> -3.432  5% -> -2.862  10% -> -2.567 |
| Gas | -5.183 | 0.000009 | 1% -> -3.432  5% -> -2.862  10% -> -2.567 |
| Water | -4.54 | 0.0002 | 1% -> -3.432  5% -> -2.862  10% -> -2.567 |

From the table, we can see that all the datasets for the Queen’s Building show that it is non-stationary (has trends & seasonality)

### Correlation

We need to see how the 3 datasets are correlated with each other and to achieve this we use a Pearson correlation function:

Where

r – correlation coefficient

- values of the x-variable in the data

- mean of the values of the x

– values of the y-variable in the data

– mean of the values in y

Chart, treemap chart

Description automatically generated

From the figure, all datasets are positively correlated with the electricity and water datasets having the highest correlation. Further examining the correlation, we can see that this is not the case as the values never intersect but Electricity and Gas have the same pattern

Chart, line chart

Description automatically generated

Example showing a correlation between Electricity and Water Datasets

Chart

Description automatically generated

Example showing a correlation between Electricity and Gas Datasets

### Feature Importance

XGBoost has a feature importance function which determines how each feature contributes to the selected target variable during regression.

Target Variable – Electricity

Chart, bar chart

Description automatically generated

The two most important features are Gas and Water which is expected, and temperature plays an equal role in electricity usage.

## Hugh Aston Building Dataset

### Preprocess/Cleaning

A black screen with white text

Description automatically generated with low confidence

Hugh Aston Electricity Reading Before Cleaning

A picture containing table

Description automatically generated

Hugh Aston Electricity Reading After Cleaning

### Visualize

Graphical user interface

Description automatically generated

Electricity Monthly Dataset

We can see from the figure above that there is a clear trend occurring. There was a downward trend happening from 2013 before picking back up in 2015 and maintaining that momentum till 2019 when it starts to drop again which might be an indicator of better energy-saving tactics employed in the building.

### Check for stationarity

|  |  |  |  |
| --- | --- | --- | --- |
|  | Test Statistic | P-Value | Critical Values |
| Electricity | -6.193937 | 0.000000 | 1% -> -3.432  5% -> -2.862  10% -> -2.567 |
| Gas | -5.006579 | 0.000022 | 1% -> -3.432  5% -> -2.862  10% -> -2.567 |
| Water | -6.168843 | 0.000000 | 1% -> -3.432  5% -> -2.862  10% -> -2.567 |

From the table, we can see that all the datasets for the Hugh Aston Building show that it is non-stationary (has trends & seasonality)

### Correlation

Chart, treemap chart

Description automatically generated

We see the same pattern occurring also within the Hugh Aston dataset as the Queen’s Building dataset where the highly correlated features do not show any correlation visually.

Chart

Description automatically generated

Electricity and Gas Correlation

As the electricity consumption rises so does the gas consumption which is expected during warm and cold months.

Chart, histogram

Description automatically generated

Electricity and Water Correlation

### Feature Importance

Chart, bar chart

Description automatically generated

By applying XGBoost regressors, the important feature of the dataset is shown above. Similarly, with the previous dataset, Water and Gas features show to be greatly important to electricity consumption. The only difference is that the air pressure shows to be more important than the temperature which would need further examination.

## Gateway House Building Dataset

### Preprocess/Cleaning

A black screen with white text

Description automatically generated with low confidence

Gateway House Electricity Dataset before cleaning

A picture containing table

Description automatically generated

Gateway House Electricity Dataset after cleaning

### Visualize

A picture containing graphical user interface

Description automatically generated

Electricity Monthly Dataset

There is a very distinct downward trend occurring which shows that the Gateway House electricity consumption rates are increasingly getting better through excellent energy-saving tactics. This shows that even without the pandemic happening the rate of consumption will always improve.

### Check for stationarity

|  |  |  |  |
| --- | --- | --- | --- |
|  | Test Statistic | P-Value | Critical Values |
| Electricity | -3.446605 | 0.009469 | 1% -> -3.432  5% -> -2.862  10% -> -2.567 |
| Gas | -4.618122 | 0.000120 | 1% -> -3.432  5% -> -2.862  10% -> -2.567 |
| Water | -5.722182 | 0.000001 | 1% -> -3.432  5% -> -2.862  10% -> -2.567 |

From the table, we can see that all the datasets for the Gateway House Building show that it is non-stationary (has trends & seasonality)

### Correlation

Chart, treemap chart

Description automatically generated

We see the same pattern occurring also within the Hugh Aston dataset as the Queen’s Building dataset where the highly correlated features do not show any correlation visually.

Chart, bar chart

Description automatically generated

At the start of the dataset, the electricity and gas features seem to struggle but they started exhibiting normal behaviours afterwards till 2020 when the sudden drop in usage happened.

Chart, histogram

Description automatically generated

### Feature Importance

Chart, bar chart

Description automatically generated

By applying XGBoost regressors, the important feature of the dataset is shown above. Similarly, with the previous dataset, Water and Gas features show to be greatly important to electricity consumption. This is extremely like the Queen’s Building dataset and shows that temperature impacts the rate of consumption.

## Weather Dataset

The analysis of the weather dataset has been broken into 3 sections; Preprocessing, Visualization and Feature Selection. The last section will be the final data which will then be added to our building data for training as a regressor.

### Preprocessing

#### View Data

Graphical user interface, text

Description automatically generated

From the above figure, we can see that there are a lot of features that could be used and some showing null values. The next step is to remove any null and incomplete data.

#### Remove null value

This can be achieved by using the built-in dropna() function that is available in python and specifying the method of removal which in this case would be ‘any’ null values found in the columns. The resulting data then becomes:

Text

Description automatically generated with medium confidence

#### Encode Categorical features

There are still some features to be discussed which are weather id, main, description and icon. These were found to be categorical data and so will be encoded. One of the features (weather\_id) had no sensible data that could be useful and was dropped from the table.

A screenshot of a computer screen

Description automatically generated with medium confidenceA screenshot of a computer

Description automatically generated with low confidence

#### Remove unnecessary features

This is the final stage of the weather dataset preprocess, and we looked at the features that bore no significance to the data. By using the drop function, we selected the following features:

* Latitude and Longitude: these are just coordinates and do not serve well as regressors since they cannot be categorically encoded.
* dt: this is the time of data calculation which is a duplicate of the dt\_iso column that serves as the index.
* timezone: these are just shifts in seconds so there is no significance
* city\_name: this is a weather dataset of Leicester.
* clouds\_all: shows the percentage of cloudiness
* wind\_deg: wind direction does not affect building energy consumption, so it was dropped.

Our final dataset before proceeding to the last step is now:

A screen shot of a computer

Description automatically generated with low confidence

### Visualization

We now must calculate the correlation between the features and plot it as a heatmap to show the feature importance. From the figure below we can see that there is a negative correlation between the humidity and temperature features. The features hardly intersect as shown below. We can see a high correlation between the dew point and temperature features. As the temperature rises, so does the dewpoint, making it similar

Chart, waterfall chart

Description automatically generated

Heatmap between features in Weather Dataset

Timeline

Description automatically generated with low confidence

Example showing the negative correlation between Temperature and Humidity

Graphical user interface, chart, line chart

Description automatically generated

Positive correlation between Temperature and Dewpoint

Below shows the final dataframe with the features and respective attributes. The count shows the total length of the dataset, mean, maximum value and the rest of the variables are also given. As shown the values are all in different ranges which would make training less accurate and difficult.

Graphical user interface, text

Description automatically generated

### Feature Selection

For this project, the temp, dew\_point, pressure, humidity, and wind\_speed was chosen as the features that would be used in training and evaluating the algorithms. The decision on these features as they have the most relevance when putting energy consumption into consideration. For example, during cold days electricity usage is expected to reduce and vice versa.

## Normalization

Feature scaling is the process of converting the values of different numeric objects to be within similar ranges. Scaling is used to prevent overfitting and biased results of supervised learning models. For example, if a model uses linear regression and the features are not scaled, some features end up having a higher impact than others, affecting the predictive performance. Therefore, it is important to scale your features. The only algorithms that do not need scaling are random forests and Decision trees since they are scaling invariant.

What is normalization? This refers to scaling features into specified ranges [0,1] in the case of a min-max scaler. Normalization is important when the data is needed in bounded intervals. Below is a formula for normalizing based on the min-max scaler:

Where:

– represents the values

– minimum values

– maximum values

### Min-Max Scaler

Min-Max Scaler is a normalisation class from the sklearn.preprocessing package. The Min-Max Scaler estimator will fit on the training data set when normalising the training and test data sets, and the same estimator will be used to transform both the training and test data sets.

Text

Description automatically generated

### Standard Scaler

Unlike the min-max scaler that normalises the values around a certain range, the feature columns are centred at a mean of 0 with a standard deviation of 1 using the standardisation approach, giving them the same properties as a typical normal distribution. It keeps relevant information about outliers and makes the algorithm less sensitive to them. The formula for standardization is:

Where:

– represents the values

– mean of values

– standard deviation of values

Text

Description automatically generated

# Experiments and Results

## Queen’s Building

### LSTM

### BLSTM

### XGBoost

### LightGBM

### Prophet

## Hugh Aston

### LSTM

### BLSTM

### XGBoost

### LightGBM

### Prophet

## Gateway House

### LSTM

### BLSTM

### XGBoost

### LightGBM

### Prophet

# Conclusion

# References