Forecasting Residential Electricity Load Demand using Machine Learning

MSc Research Project
Data Analytics

Arun Panigrahi
Student ID: x18186840

School of Computing
National College of Ireland

Supervisor: Dr. Manaz Kaleel

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Arun Kumar Panigrahi
Student ID:	x18186840
Programme:	M.Sc.-Data Analytics
Year:	2020
Module:	Research Project
Supervisor:	Dr. Manaz Kaleel
Submission Due Date:	28/09/2020
Project Title:	Forecasting Residential Electricity Load Demand using Machine Learning
Word Count:	7182
Page Count:	24

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Arun Kumar Panigrahi
Date:	27th September 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Contents

Forecasting Residential Electricity Load Demand using Machine Learning

Arun Kumar Panigrahi
x18186840

Abstract

Since the emergence of different forms of sophisticated home appliances and smart home devices globally, the demand for residential energy is rapidly growing. This growing demand has led to create energy sustainability issues, which have been identified as one of the major concerns in the recent times as more consumers need more energy. Therefore, the forecasting of demand for electricity plays a crucial role in maintaining an equilibrium between the consumers demand and energy generated by the energy producing companies. In recent years, numerous machine learning algorithms have been employed to forecast the consumers electricity load demand. This study has been carried out as a comparative analysis of four different machine learning algorithms, Autoregressive Integrated Moving Average with Explanatory Variable (ARIMAX), Seasonal Autoregressive Integrated Moving Average with Explanatory Variable (SARIMAX), PROPHET and Long-Short Term Memory (LSTM) for a group of residential customer present at London. It is observed that the efficiency of the LSTM model is considerably greater than all the other implemented models.

Keywords - Residential Energy, Forecasting, Electricity, Energy Demand, ARIMAX, SARIMAX, PROPHET, LSTM

1 Introduction

The pace of development of countries worldwide is now progressively increasing and seems to be inevitable. The increasing trend of human population clearly demonstrates that demand for housing and the growth of infrastructure is going to steadily increase. This rapid development of residential and business areas, particularly through industrial growth, contributes in increasing rates of energy consumption. In addition to all these advancements, surplus energy supply is essential to drive global demand while maintaining the environment safety (Afshar et al.; 2012). According to the 2017 World Energy Outlook reported by IEA[1] (International Energy Agency), global demand for energy is projected to increase by 1.0 % CAGR (Compound annual growth rate) in the period 2016-2040.

A significant number of studies have been compiled in recent years on exact load forecasting, owing to their effect on activity of energy generation and supply systems as

[1]https://www.iea.org/reports/world-energy-outlook-2017

well as the economies. As a result, these studies have positively impacted to generate technically advanced and powerful load forecasting models at independent infrastructure levels. The approaches used to forecast energy consumption in any forms are usually categorized in two segments: engineering (physical) approach and techniques based on customer load consumption data. Engineering approaches utilize arithmetic equations to describe the physical structures and thermal efficiency of buildings. Moreover, engineering approaches require comprehensive information on different building specifications, which are not often accessible, and a high degree of experience is required to conduct this kind of expensive and rigorous computation. In comparison, data driven methodologies do not require detailed information about the structure of the building and rather learn from historic or real-time data of consumers electricity usage (Zhao and Magoulès; 2012) (Pérez-Lombard et al.; 2008).

The forecasting of electricity load refers to calculation of customers energy demand in the future. Load forecasting enable suppliers to manage supply and demand along with maintaining the stability of energy supplies even if there is a power shortage (Weron; 2014). In this instance the energy performance of the housing sector is an important objective and potential resources for policymakers. As around one fifth of the global demand for electricity arises from the residential sector – out of the demands of heating, cooling, and lighting homes. Therefore, electricity demand forecasting of the residential sector is a vital for energy management and power supplying companies (Almeshaiei and Soltan; 2011).

Load forecasting is applied in a wide variety of time horizons for various strategic objectives: (i) Short-term load forecasting (a few hours to 1 day ahead), usually considered for supply and demand management, (ii) Medium-term load forecasting (few days to 1 year ahead), typically used for outage and other maintenance related operations. (iii) Long-term load forecasting (a year and ahead) commonly used to execute energy infrastructure advancement. Load forecasting is often performed at various aggregate levels in areas with specific regional dimensions such as countries, counties, community, or houses. The forecasting task is particularly complicated with respect to lower rates of aggregation such as the building level as many fluctuating variables influence the energy usage of a building at various stages, such as environmental conditions, construction locations, heating and ventilation systems as well as the consumption behaviour of inhabitants. Moreover, a precise demand forecast will contribute to a substantial reduction in operational and repair expenses, enhance the stability of the power supply and distribution network along with optimal potential planning and management (Mocanu et al.; 2014) (Pirbazari et al.; 2020).

Over the past couple of years, smart meters are being implemented exponentially across the globe. By end of 2018, nearly 11 million smart meters were deployed in United Kingdom (UK) by the energy suppliers, and about 90 % of such meters have been implemented for residential consumers. These Smart meter-generates high resolution data, which provide suppliers with numerous control functions, including tracking of service quality and power loss detection. This also open various doors for power demand monitoring, such as highly accurate load forecasting at lower aggregation rates (Ushakova and Murcio; 2018).

According to different literature, the most conventional Machine Learning and Deep Learning based techniques formulated specifically for load forecasting needs an ample amount of historical load data for training and testing. In case of residential customers, mostly the historical energy consumption data of the building is accessible and are typically used for model training and testing. Additionally, many researchers utilize other external information and create multivariate models focused on consumer behaviour and environmental factors etc. to enhance the prediction precision.

This research explores various Machine Learning and Deep Learning based approaches for electricity load forecasting, while as a state-of-art a sample of smart meter data of residential building at London has been considered. In this regard, four different models to forecast the regular average housing load consumption have been established and their predictive precision and generalization capability have been evaluated in the specific configurations. The models used in this study were identified from the extensively used in general forecasting approaches such as the ARIMAX, PROPHET, SARIMAX and LSTM standard-architectural versions. These models are being trained and tested on the daily average electricity consumption data of numerous domestic customers. In addition, models are evaluated for their predictability to the scale of training data and to the amount of input variables and a systematic assessment of forecasts outcomes is performed. Moreover, to improve the generalization capabilities and enhance model robustness, the models developed are expected to learn from the built-in knowledge in time series data specific load profiles.

1.1 Research Question and Objective

How well can Autoregressive Integrated Moving Average with Explanatory Variable (ARIMAX), Seasonal Autoregressive Integrated Moving Average with Explanatory Variables (SARIMAX), PROPHET and Long Short-Term Memory (LSTM) algorithms forecast residential electricity load demand for a sample of residential units in London?

The objective of this research project is: Forecasting the domestic electricity load demand for a sample of residential units in London, using machine learning and neural networks. In particular, the research project will concentrate on the below mentioned objectives:

- Forecasting electricity load demand by considering attributes such as, past load consumption and meteorological conditions.
- Performance analysis of employed machine learning and Neural Network models.
- Comparison of the forecasting outcomes to find out the best fit model.

2 Literature Review

2.1 Heterogeneity of European Residential energy demands

The power production systems have moved from structures with a specific number of major participants with conventional power stations, to systems that have more smaller companies involved. Almost all European countries are going through this constant transformation in power generation and distribution sector. The growing share of renewable sources of energy contributes in more decentralized and continuously changing system for generation of electricity. Additionally, exponential growth of market trends has been building fresh challenges in the balance between supply and demand. (Hayn et al.; 2014)

Although residential sector is accountable for 29 % of overall energy consumption in Europe, their demand patterns are yet to be identified precisely. (Hayn et al.; 2014) proposed strategies consisting of household clustering and segmentation that primarily emphasized on electrical appliances as well as emerging power and heat supply solutions. In this study a thorough evaluation of the social and demographic effect on residential electricity load demand was carried out. Summarizing the outcome of the approaches, that were utilized to explain energy demand in residential sector, it was clear that household electricity demand forecast is a complex operation. The energy consumed by any house is the amount of the energy utilized by all the household appliances. Although the amount and type of device are directly related to income and size of household, however, the habits of usage are rather dependent on the state of occupations of the house owners. Ultimately, it was identified that, four different attributes such as – lifestyles, socio-demographical influences, household appliances and power generation technicalities have a great deal of impact on household energy demand trends and overall energy usage.

(Drysdale et al.; 2015) study stated that the gross load demands of residential consumer in United Kingdom is expected to raise up to 68.2-Terawatt Hour (TWh) annually by 2030. In yet another study by (Ramírez-Mendiola et al.; 2017) it was reported that the average domestic energy demand figures of individual residential unit range from 2000 to 7200 kilowatt-hour (kWh) per year referring to statistics from the Office of Gas and Electricity Markets (OFGEM) UK. This volatility in energy usages is therefore correlated with several causes, such as variations in household requirements and the effects of environmental conditions. A separate study by (Meier and Rehdanz; 2010) revealed the deciding factors for the heating actions of the residential units in UK. It was discovered that socio-economic factors such as household earnings and expenditures plays a significant role in the reported variations of heating and cooling expense of buildings alongside the sources of energy used.

2.2 Environmental impact on electricity consumption

The correlation between environment and energy consumed by residential electric appliances has been extensively analysed till date, but it remains ambiguous. The unforeseen consequences of climate change, which are assumed to intensify climate and environmental disasters, exacerbate the uncertainties. (Nateghi and Mukherjee; 2017) presented a non-linear statistical learning technique to develop predictive models that explains the correlation between weather, climate, and electricity consumption of domestic and commercial consumers. The findings of the study clearly indicated that climatic condition plays a significant role in influencing the energy consumption curves of both the residential and commercial electricity consumers. It was observed that the monthly energy usage throughout the residential sector declines initially with a rise in Mean Dew Point Temperature (MDPT).

While the energy market dynamics and volatility are constantly evolving, (Staffell; 2017) designed an open framework to evaluate environmental impacts on power supply and demand. The outcome of the analysis stated that the UK power grid is shifting rapidly in unparalleled ways, with maximum power demand reaching 70 Gigawatt while j respect to space heating and cooling. A statistical analysis was carried out for each household with variations from the hourly average energy usage at each temperature level. It was concluded that customer behaviour with respect to the climatic condition illustrate the load curve for the residential energy usage.

2.3 Electricity Load Prediction Using Machine Learning and Deep Learning

A precise and reliable electricity consumption forecasting is a key factor in sustainable energy management system. (Bouktif et al.; 2018) proposed a univariate demand-side load forecasting model based on LSTM-RNN, that accurately estimated for both the short-term and medium-term electricity load requirement. In order to perform a comparative analysis, the best performing model was used as a benchmark out of seven different forecasting techniques utilized in this study, which illustrated a range of widely employed machine learning models in the term of energy load forecasting. It was observed that LSTM models with optimum time-lagged features captured all the aspects of the complex time series and produced very accurate results.

An accurate forecasting of electricity load is the key role in operation and management of power grid networks, to maintain an optimum equilibrium between demand and supply, along with minimization the costs of production. (Guo et al.; 2018) employed a deep feed-forward network in an analysis to predict short-term energy demands. Initially the customers consumption patterns of electricity load are analysed on the basis of monthly, weekly load consumed and along with the trends of temperature. Furthermore, a method for probability density forecasting is proposed relying on deep learning, regression models and an estimate of kernel density. Finally, the findings were compared against some of the common machine learning models such gradient boosting and random forest. Experimental findings demonstrated that deep leaning technique had a lower forecasting error and better performance than the machine learning models such as random forest and gradient enhancement approaches.

The Demand forecasting for a single residential consumer is a difficult job because of the volatility involved at the granular scale. (Chinnathambi et al.; 2018) analysed the advantage of recurrent neural networks over conventional methodologies with smart metering data sets. The preliminary findings of the study revealed that LSTM dependent recurrent neural network (RNN) was more effective for a single consumer with 1-minute resolution dependent on one-year historical data sets compared with a simple RNN and gated recurrent unit RNN. (Wang et al.; 2019) analysed the usage of Automated Machine Learning techniques over multiple real-world data sets to conduct electricity load forecasting operation. The finding of the study demonstrated clearly that Automated Machine Learning can significantly decrease effort in the development of effective energy predictive models and enhance the outcome accuracy while obtaining ideal energy consumption forecasting work.

2.4 Adaptive load forecasting

It is well-known that electrical load forecasts are made on different time scales from short, medium, and long-term load forecasts. (Santika et al.; 2017) developed a mechanism focused on the adaptive neuro-fuzzy inference system (ANFIS) that involves no specification by subjective means of individual initial parameter for membership function. Adaptive neuro-fuzzy inference process is a framework that integrates fuzzy logic with adaptive neuro model and is applied in the time-series prediction problem. The complete mechanism of the model can described as follows, Firstly, the configuration of ANFIS is determined by subtractive categorisation: secondly, the values and consistent parameters of ANFIS is defined by means of a hybrid algorithm; finally, certain factors influencing future day-to-day electricity, such as weather and population, are considered as input parameter from ANFIS to forecast daily electricity load requirement. The statistically significant results exhibited that ANFIS had effective predictability with a fast running time and therefore evaluated that the approach is feasible.

One of the common challenges in load forecasting operation is to customize the model fit in the load prediction according to the requirement of substations along with minimizing IT resources and regenerate these statistical models to wide variety of data sources. The prime objective of a study proposed by (Wee and Nayak; 2019) was to identify a novel framework that respond to the load demands by integrating reinforcement learning with load prediction on presently available database technology. Hence, the suggested approach evaluates and utilizes better models to cope with the varied levels of accuracy, than these models are re-calibrate in an iterative way comparing the forecast outcome against the actual results. Empirical findings of the study demonstrated that the better operating platform enhances the process by boosting computational capacity, storage efficiency and higher performance. (Zunic et al.; 2020) established an adaptive conditional density estimation (ACDE) method based on kernel density estimation (KDE) in order to optimize the load forecasting accuracy. The suggested framework is builds upon ACDE to extract the total electricity demand from the temperature dependent element, such as heating and air conditioning usage. On validating the output of the ACDE framework by a comparison to the adaptive filter models of multi-layer perceptron and recursive least-squares revealed that this technique is best suited for multi-stage forecasting. Moreover, the comparison outcomes from real data reveal that the ACDE model could increase the accuracy of the predicted load profiles across the entire horizon versus the other models.

3 Methodology

This section describes the methodology used alongside the project flow diagram which has been formulated for the project execution. The techniques in the development of the project were followed in accordance with the Knowledge Discovery in Databases (KDD) methodology represented in figure 1.

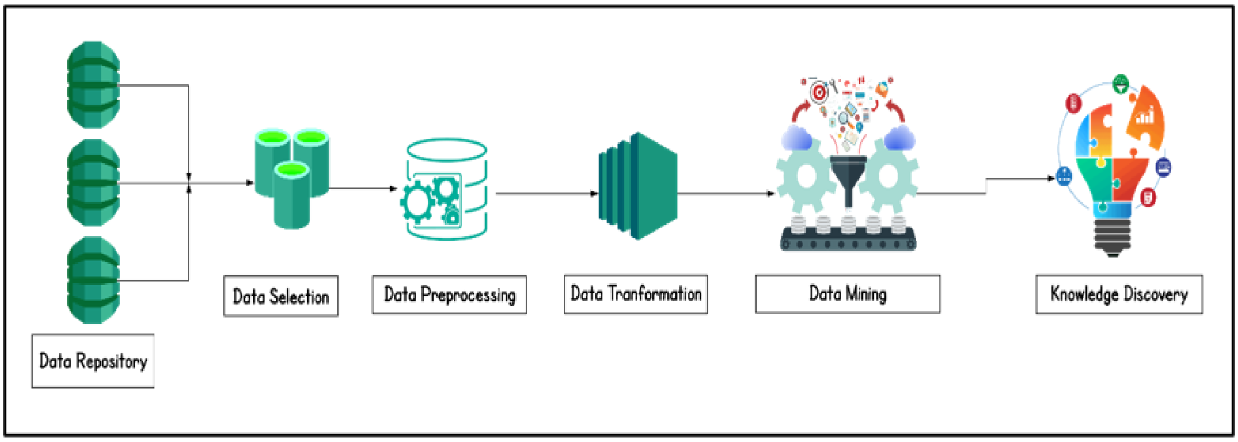

Figure 1: Representation of the relationship among different phases of KDD methodology

3.1 Data Selection

In this phase of the project, the data required for the Knowledge Discovery processes is determined. This involves the exploration of whatever data is available, the collection of critical information, and the aggregation of all knowledge discovery information into a common set of characteristics that would be used for the process. This method is very crucial since the data mining framework is developed in order to discovered knowledge from the available data set and the evidence base for model building (Xydas et al.; 2013).

The data set utilized for this project has been extracted from the online repository "Kaggle.com", these data-sets consisted of multiple files of time series data such as electricity load consumption from various smart meters of residential customers present at London and daily weather reports.

3.2 Data Pre-Processing

This stage is one of the most significant steps in the process since it involves the cleaning of data by removal of null values and exploring the data in order to gain insights for the advancement into the next stages of the analysis (Xydas et al.; 2013).

The electricity smart meter data set comprised of attributes such as, average electricity consumed per household, MeterID of the House and Date. After testing for the null value, it was observed that some of the instances composed of missing values. Therefore, the target variable was first transformed into a time-series object using the ts() function of R Programming Language, to handle the missing values in the data set. Then the linear interpolation technique was employed to impute the missing values. Similarly, the weather data-set curated form "Kaggle.com" which includes several attributes like "maximum temperature", "minimum temperature", "wind speed", "UV rating", etc. It was observed that the data set does not have any missing values.

3.3 Data Transformation

Data transformation involves a range of feature engineering techniques that are utilized to further support the relevant assumption and to enhance the features of data set for data modelling (Xydas et al.; 2013). Therefore, it is always preferable to appropriately transform time series data to effectively model it in the supervised learning process.

In this project to accomplish this challenge, various new features were developed from the electricity and weather time series-datasets, For example the timedata() and Lubridate() library in R programming language have been applied to retrieve additional features from the time stamp attributes present in both data sets (Weather and electricity consumption).

3.4 Data Mining

The implementation of appropriate data mining algorithm is essential in order to achieve the research objective. The key factor for selecting an appropriate algorithm depends upon the form of data set used in the analysis. As the data set considered in this analysis comprises of various data captured over a duration of time, therefore time series models was determined for the implementation (Mariscal et al.; 2010). The machine learning models mentioned below had been considered for this research,

3.4.1 ARIMAX

The regular ARIMA (Autoregressive Integrated Moving Average) model can render predictions dependent on the predictor values of the past. ARIMAX model can be explained as the extended version of ARIMA. It can be defined as autoregressive moving average with other independent or predictor variable. It is also termed as dynamic regression model as it is equivalent to a multidimensional regression model as well as it permits the auto-correlation present in regression residues to enhance the forecasting accuracy. This technique is appropriate for the forecasting of stationary or non-Stationary data and for multivariate data trends of any kind (Ulyah et al.; 2019). The generic notation of an ARIMAX model is represented as arima(p,d,q) along with xreg where,

- xreg = predictor/independent external variable
- p = lag or number of lag present in the observation,
- d = degree of differencing and,
- q = order of moving average.

The significance of p, d and q is determined according to the Box-Jenkins Method. Also, the Augmented Dickey Fuller test (ADF) assists in identifying the significant value of p that is the lag order. Similarly, an Partial Auto-Correlation Function (PACF) plot demonstrates the value of q, by notifying the lag that is the last significant lag in the plot (Ulyah et al.; 2019).

3.4.2 SARIMAX

The SARIMAX model is an improved version of the Seasonal Autoregressive Integrated Moving Average (SARIMA) model, in which the predictive efficiency is enhanced by the integration of exogenous variable. The exogenous parameters can be defined as the variables that can affect the model outcome but does not get effect itself. The climatic condition such as temperature can be considered an exogenous feature in the context of a building an energy demand forecasting model. The generic notation of an SARIMAX model is represented as SARIMAX (p, d, q) (P, D, Q)s where,

- p = order of the Autoregressive term of the predicting variable.
- d = order of differencing which is required to make the data stationary for the predicting variable.
- q = order of Moving Average term of the predicting variable.

  Similarly,

- P = order of seasonal autoregressive term.
- D = order of seasonal differencing which is required to make the data stationary.
- Q = order of seasonal moving average.
- s = number of seasonal intervals

The Augmented Dickey Fuller test (ADF) and Partial Auto-Correlation Function (PACF) test of the dependent and the exogenous variables, helps in identification of p, d, q, P, D, Q. Additionally the s is identified by seasonal difference present in the time-frame of the data. (Elamin and Fukushige; 2018)

3.4.3 Prophet

The Facebook core data science team has developed a model known as Prophet that can be used for daily, weekly, and yearly non-linear trend forecasting. The model is adaptive in nature and uses a modular regression method that enables the selection of resources relevant to a forecasting problem and makes adjustment wherever applicable while also functioning well with default parameters. It also provides a forecast analysis and evaluating system that allows users to modify and improve the forecasting outcome by continuous improvements. It fits well for time series data that have strong seasonal factors and historic seasonal data. Prophet is robust in the processing of data with missing values, trend shifting values and handles the outliers present in the data excellently (Asha et al.; 2019).

In this project, the Facebook prophet algorithm is implemented to model the dynamics of regular residential electricity load with the goal of forecasting the potential demand for electricity.

3.4.4 LSTM

The Long Short-Term Memory (LSTM) is a form of Recurrent neural network (RNN) algorithm that can recall the values from the previous phases and keep as a reference in order to forecast the future values. LSTM is a series of cells or system components that accumulate and save the data feeds. These cells represent a transmission line that transfers the data from the past through one node to another and stores them for the present. In addition, through the use of certain gates within each cell, data can be discarded, extracted, or added in the next cells. Furthermore, cells can either allow data to be passed through or disposed of using the gates depending on the sigmoid functions of neural network layer. These sigmoid layer yields numbers within the ranges of 0 and 1, which represents the distribution of data in each segment of the cell. (Le et al.; 2019)

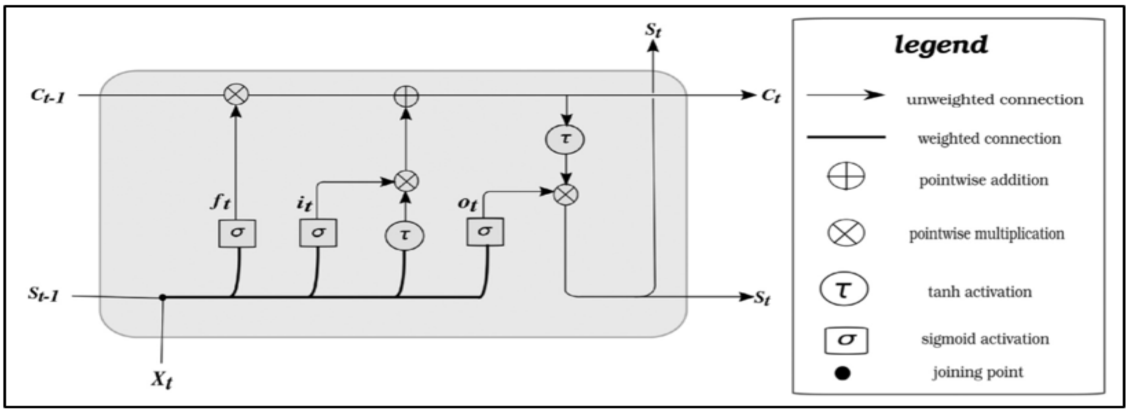

Figure 2: LSTM Hidden Layer Architecture

(Siami-Namini et al.; 2018) in a study stated that the typical LSTM model comprises of one hidden layer which obtains the input value from three separate points which include, present value as X(t), output from the hidden layer as s(t-1) and the unit state as c(t-1). Also, it consists of three different kind of gates with the purpose of regulating the condition of each cell and those are,

- Forget Gate (f_t) – It generates a number between 0 and 1, in which 1 indicates "store the complete value" and 0 represents "to forget it absolutely."
- input Gate (i_t) – It assists in selecting fresh data that is required to be saved in the cell. Initially a sigmoid layer selects which values would be modified and finally, a tanh layer generates a vector of new values which can be added to the state.
- Output Gate (o_t) – It determines what is supposed to be the output of each cell. These output values are dependent on the cell status as well as the processed and freshly added data.

3.5 Evaluation

The computation of model performance and efficiency is one of the critical phase of the analysis, as it helps to analyse the model developed and determine its effectiveness to address the specified problems and goals. In this phase, the knowledge gained following the application of four time series forecasting algorithms is assessed to gain clear insight further into intended outcome. This understanding is then visualized to provide a quick, practical insights into the electricity demand details in front of end users. The evaluation metrics used in this analysis are Mean absolute error (MAE), Mean absolute percentage error (MAPE), Root-mean-square error (RSME) and mean squared error (MSE). (Wang et al.; 2009)

4 Design Specification

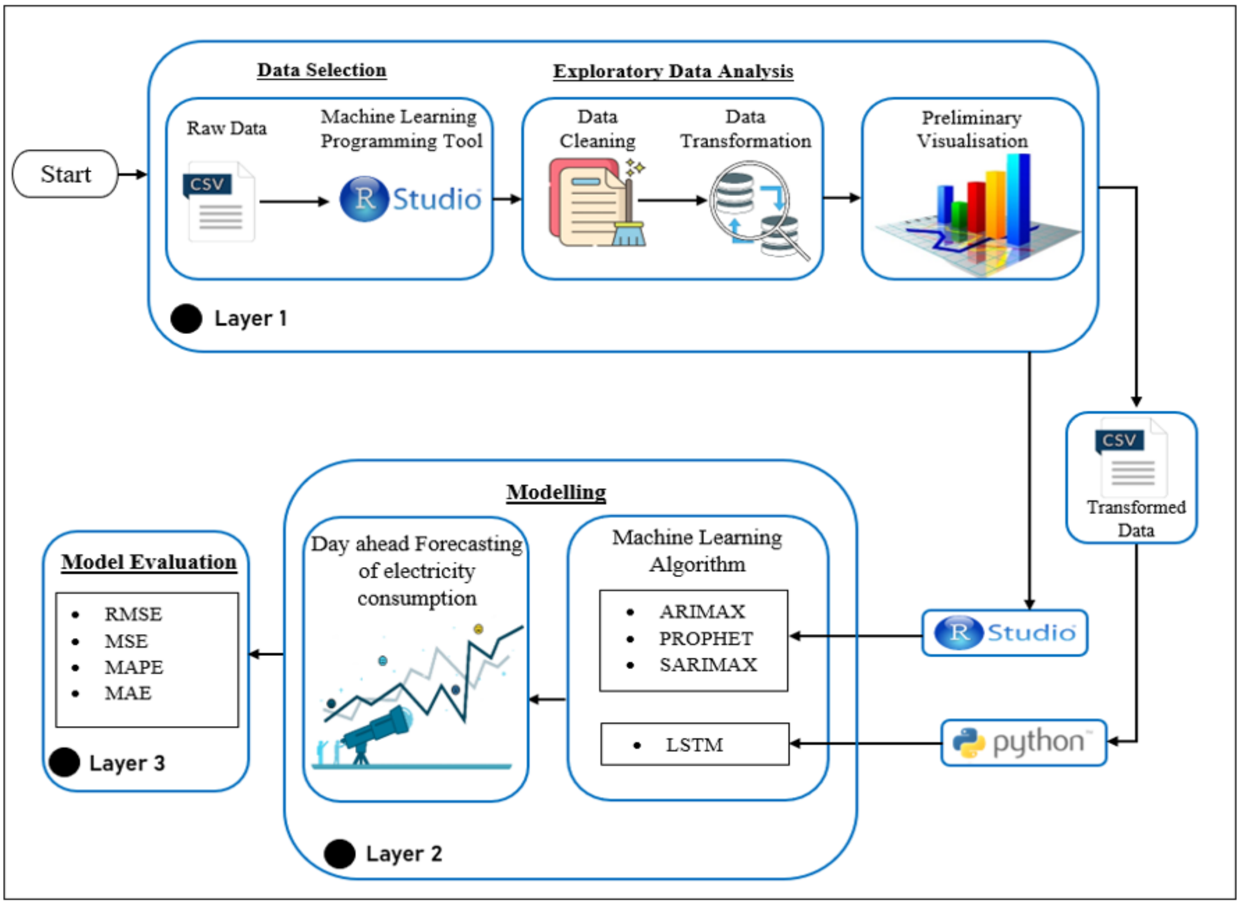

Figure 3: Implementation and Solution Design

The entire project execution procedure is separated in three different layers. In layer-1 the electricity usage data of residential buildings in London was retrieved from the data repositories of kaggle.com. Subsequently, data pre-processing was performed using R programming language. Data modeling and visualisation of the model outcome was carried out in the layer-2. Finally in the Layer-3 the overall performance of the model and its outcomes were analyzed by various statistical computation technique, which leads to the selection of model with best performance and most accurate forecasting outcome.

5 Implementation

This phase provides a comprehensive overview of the different steps involved in developing the ARIMAX, Prophet, SARIMAX and LSTM algorithms to forecast residential power demand.

5.1 Environmental Setup

The project has been carried out on R and Python programming language. The initial data pre-processing and implementation of ARIMAX, SARIMAX, and Prophet models has been done using various statistical packages in R Studio. Additionally, LSTM model is implemented using Keras and Tensorflow libraries of python programming in Jupyter notebook. The benefit of utilizing Keras and Tensorflow library is that it is convenient to execute and well suitable for the neural networks algorithms.

5.2 Selection of Data

The data curated from kaggle.com[2] data repository consisted of electricity load consumption reading from smart meters for a sample of 5566 London households that took part in a project led by UK Power network[3] between November 2011 and February 2014. Additionally, the weather data of London area between November 2011 and February 2014 was also retrieved. The electricity load consumption dataset composed of 9 attributes and 3,510,433 instances, while 882 instances and 32 attributes were found in the London weather data-set. These data-sets as csv file were converted into data-frames upon uploading into the R-Studio for further processing.

	Category	**Data Source**	**Description**	**Features**	**Target Variables**
1	Weather Data	Kaggle.com / darksky.net	Meteorological data including detailed weather information of London curated from darksky.net	Minimum daily Temperature, Maximum daily Temperature, wind bearing, visibility and Humidity	Average daily Temperature (K)
2	Electricity Load data	kagglecom / https://data.london.gov .uk/	Electricity consumption readings from Smart meters for a sample of 5,567 London Households, between November 2011 and February 2014	Time Stamps, Meter ID, Total Electricity Consumed, Average Electricity consumed per day, etc	Time Stamps, Meter ID, Average Electricity consumed per day (KWH)

Table 1: Data Set Description

[2]https://www.kaggle.com/jeanmidev/smart-meters-in-london
[3]https://data.london.gov.uk/dataset/smartmeter-energy-use-data-in-london-households

5.3 Exploratory Data Analysis

5.3.1 Data Cleaning

In this stage of the project, the prevalence of the null or NA values in the data-set has been scanned. In order to eliminate the missing values from electricity load data, The ts() function and linear interpolation technique were used. Firstly, the ts() function was used to transform the values of target variable that is "energy_sum" from numeric vector to R time series object. Next, the linear interpolation technique was used to identify and replace the missing values in a time series object. Finally, the cleaned data was eventually restored back to the data frame. Similarly, it is identified that certain attributes of the weather data-set composed of null values, however, these attributes had not been considered in this study. Thus, it is completely secure to overlook these null values and continue further in the analysis.

5.3.2 Data Transformation

Transformation of raw data into suitable format for a better understanding of machine learning algorithm is one of the most essential stage of any data mining project.

The electricity load consumption and weather detail data consisted of date attributes in order to specify the regularity of the data. Initially, timedate() and lubridate() function of the R Programming were employed to retrieve the year, month and date constraints out of the date attribute.

The electricity load data composed of average power usage of 5,566 residential households in London. Hence, in order to evaluate the average electricity consumption of over all residential building combined together, the electricity load data was grouped by date using group_by() function along with computing the mean electricity consumed by all residential households in one day.

Furthermore, in order to find the target variable that is the average mean temperature of each day, computation was performed using the mathematical formula described below. These calculated values were than added to the weather data frame as average temperature attribute. Also, the weather data attributes were grouped by date feature after computing the average daily temperature.

$$\frac{MaximumTemperature + MinimumTemperature}{2}$$

Subsequently, both these data-sets is concatenated in accordance with the date attribute using the inner_join() function. A computation of the weekly and monthly moving average of daily electricity usage and average temperature was also performed for the purposes of preliminary visualization and further analysis.

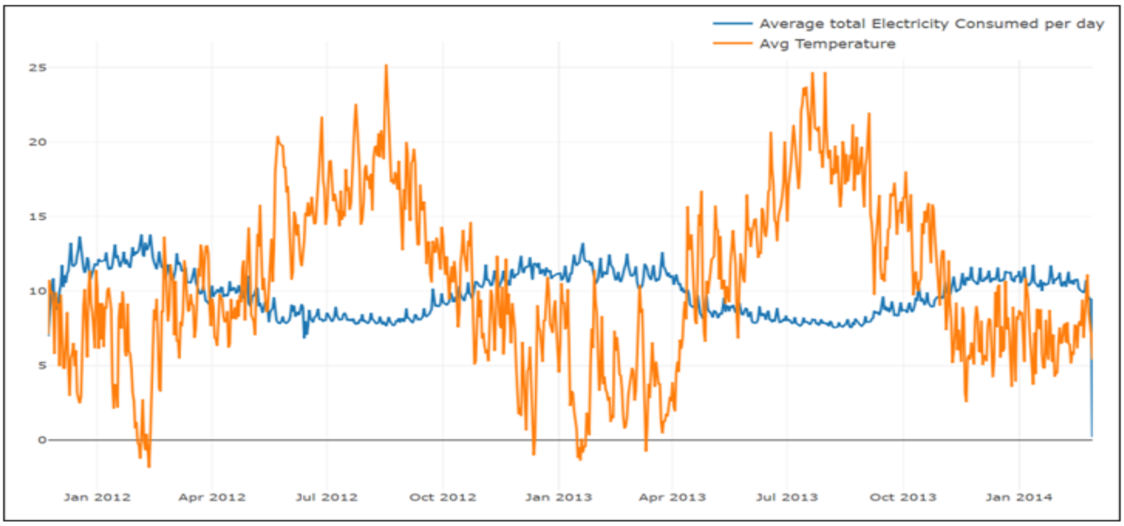

Figure 4: Average Daily Electricity Consumption V/s Average Daily Temperature

The graphs demonstrated in the Figure 4 indicates the seasonal variations in the time series data of average residential electricity consumption per day in comparison with the average daily temperature. The existence of seasonal variance in the daily electricity consumption data is evident as a slight decline in the temperature increases the electricity usage trend and vice versa.

5.4 Model Implementation and Evaluation

5.4.1 Experiment 1 - ARIMAX

In order to implement a multivariate timeseries forecasting technique, ARIMAX model has been considered. As mentioned below certain mandatory data pre-processing procedures have been performed before the application of ARIMAX model.

In this phase of the project, both the forecasting and independent variables were retrieved from the real data-set and transformed to a time series vector using ts(). Furthermore, seasonal and trend decomposition using Loess (STL()) function was used to identify and decompose the time series data. Also, seasadj() function was used to adjusted the seasonality variance in the data by removing the seasonal component.

Additionally, Augmented Dickey Fuller (ADF) test was carried out in order to verify the stationarity of the data. It was found that both the data, electricity load consumption (deseasonal_cnt) and weather data (deseasonal_cnt_wether) indicated a Lag order = 9 and p-value = 0.01, also stating the alternative hypothesis as stationary. The figure-9 indicates the ACF and PACF test result for the electricity data-set. It can be therefore assumed from the PACF plot that the Periods to lag is 7. This lag value is considered after the PACF crossed the positive confidence interval for the first time.

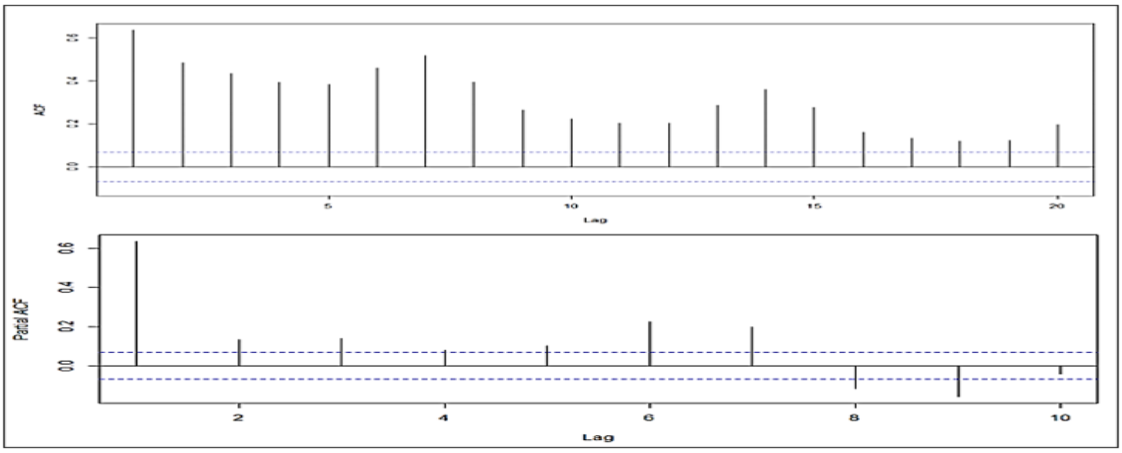

Figure 5: ACF and PACF TEST

In order to implement the model and test its accuracy, the data set is divided into training and testing set in 7:3 ratio. As represented in the equation (1), the forecasting attribute (daily average electricity load consumption) is presented as training_electricity and the external variable (daily average temperature) xreg is supplied as training_weather.

$$arima(training\_electricity, xreg = training\_weather, order = c(9, 0, 7)) \tag{1}$$

The value of p is defined as 9, which is found from the ADF test lag order value, value of d is defined as 0, because the difference of the data point has not be considered, while the value of q is defined as 7, which is assumed from the PACF plot as presented in Figure 5 .

After the model has been trained as mentioned, next predict() function has been applied to forecast future values up to 251-days. The graph present in Figure 6 indicates the real values present v/s the forecast values. The black line in the figure indicates the real value while the red line indicates the forecasting outcome. It can be clearly noticed that there are some major variation between both the real and the predicted value of electricity usage towards the end of the graph.

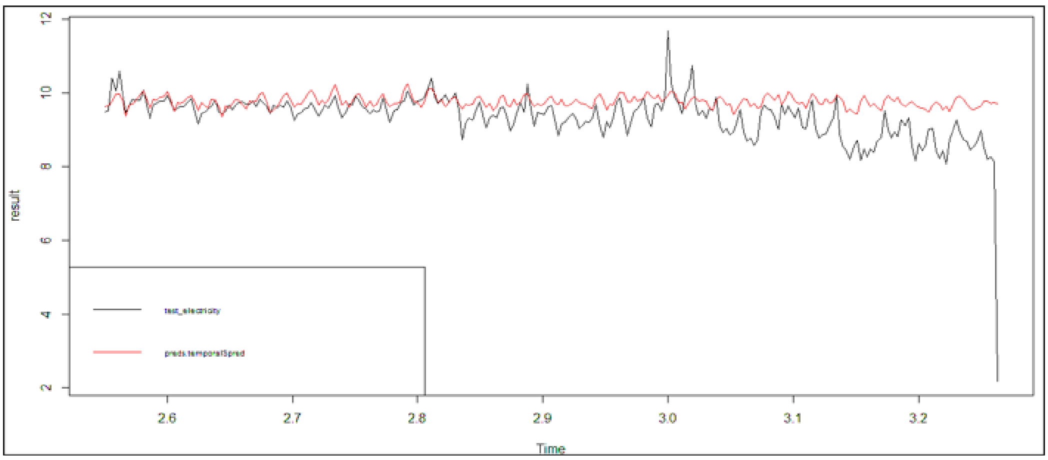

Figure 6: Real v/s Forecast Outcome - ARIMAX Model

Evaluation - In order to evaluate the efficiency of the model, multiple evaluation matrices has been used such as Root Mean Square Error (RSME), Mean Square Error (MSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) were used. As presented in the Table 2, the Mean Square Error value was calculate was 0.570 and the Root Mean Square Error was 0.755. Similarly, the Mean Absolute error was computed as 0.469 and the Mean absolute Percentage Error was computed 6.21%.

RMSE	MSE	MAE	MAPE
0.76	0.57	0.47	6.21

Table 2: Evaluation Metrics Result of ARIMAX Model

5.4.2 Experiment 2 - SARIMAX

The implementation of SARIMAX model is considered whenever the time series data possesses some kind of seasonality. One of the key benefits of the SARIMAX model is that it can cope better with the missing data in time series without impacting its outcome (Tarsitano and Amerise; 2017).

$$fit1 = arima(training\_electricity, xreg = training\_weather,$$

order=c(9,0,7), seasonal=list(order=c(9,0,1), period=3)) (2)

In the SARIMAX modelling, factor that has been identified for the ARIMAX model was kept unchanged, additionally the seasonal order feature was introduced in compliance with the SARIMAX model specification. The data ratio of 70:30 has been retained for model training and testing. As stated in equation (2) above, the model is trained using electricity load consumption data (training_electricity) and seasonal exogenous variable as the average daily temperature (training_weather). The values of model order (p,d,q) were identified from Augmented Dickey Fuller test (ADF) and Partial auto-correlation function (PACF) test of electricity consumption test data. The characteristics of the seasonal order (P,Q,D) values were determined from the ADF and PACF assessments of the daily average temperature attribute, where it is seen the Lag order = 9 and the Periods to lag is 1.

After the model is trained, predict() function was used to forecast the future values. The graph present in Figure 7 indicates the real values present v/s the forecast values. The black line in the figure indicates the real value while the red line indicates the forecasting outcome. It is clearly evident that forecasting outcome has not been able to capture all the spike present in the real test values.

Evaluation - As present in Table 3, the findings of the evaluation matrices illustrate that, the Mean Square Error (MSE) is 0.599 and Root Mean Square (RMSE) Error is 0.774, While the Mean Absolute Error (MAE) was found 0.495 and the Mean Absolute Percentage error (MAPE) is 6.50%.

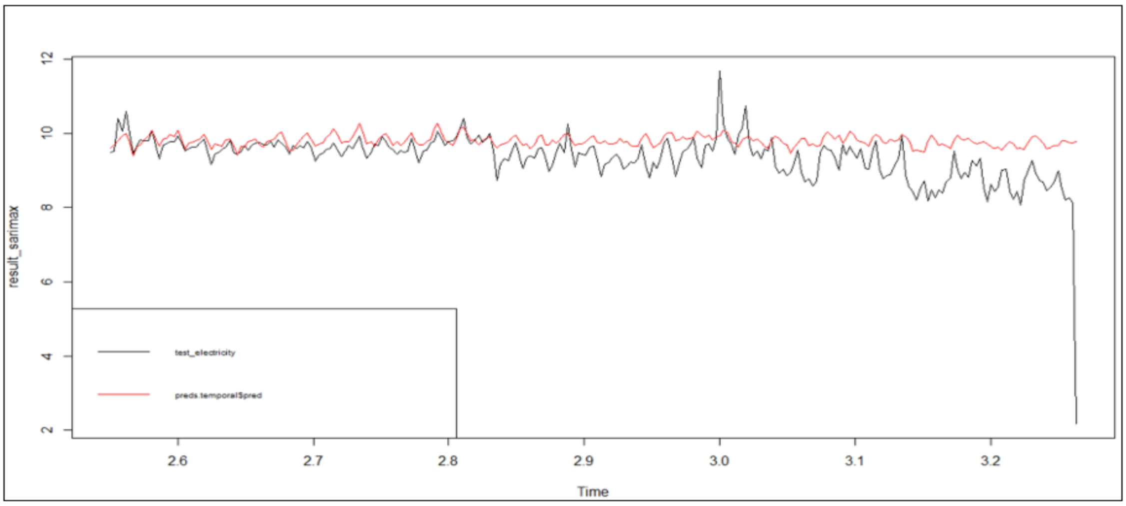

Figure 7: Real v/s Forecast Outcome - SARIMAX Model

RMSE	MSE	MAE	MAPE
0.77	0.60	0.50	6.50

Table 3: Evaluation Metrics Result of ARIMAX Model

5.4.3 Experiment 3 - PROPHET

The prophet model has been implemented by utilizing the prophet and jsonlite of R library. One of the fundamental guidelines of the prophet model is that it only recognizes the forecasting attribute as a "ds" and the forecasting time period attribute as "y" (Zunic et al.; 2020). As the average electricity load consumption (energy_sum_mean) is the forecasting parameter in this project, therefore it has been renamed to "ds", like wise the date attribute has been renamed to "y". Finally, this updated data frame was introduced to the prophet() model. Furthermore, to forecast the values 'predict' function was used. The forecast outcome was then recorded into a data frame, while in order to create the data frame 'make_future_dataframe' function was used. The following equation illustrate the operational technique of the model.

$$y(t) = g(t) + s(t) + h(t) + \epsilon t \tag{3}$$

The key three factors of Prophet model are, trend defined as "g", seasonality defined as "s" and holidays defined as "h", that allows this model decomposable. The prophet model is largely recognized because of its capacity to manage large irregular data. The key benefit of this model are its highly accurate prediction outcomes and low processing time due to simple model parameters (Zunic et al.; 2020).

Ultimately a graph was plotted with forecasting and the real value as presented in Figure 9 .The black points shown in the graph indicates the real data points and the blue demonstrates the point approximation while with the light blue area indicates the predictive interval. It noticeable from the plot that , the blue line is quite closely matches the real data points. One of most important feature of this prophet graph is its interactive visual aspect which enables the user to move the mouse over the plot to easily view the actual value and the forecast value for the time stamp.

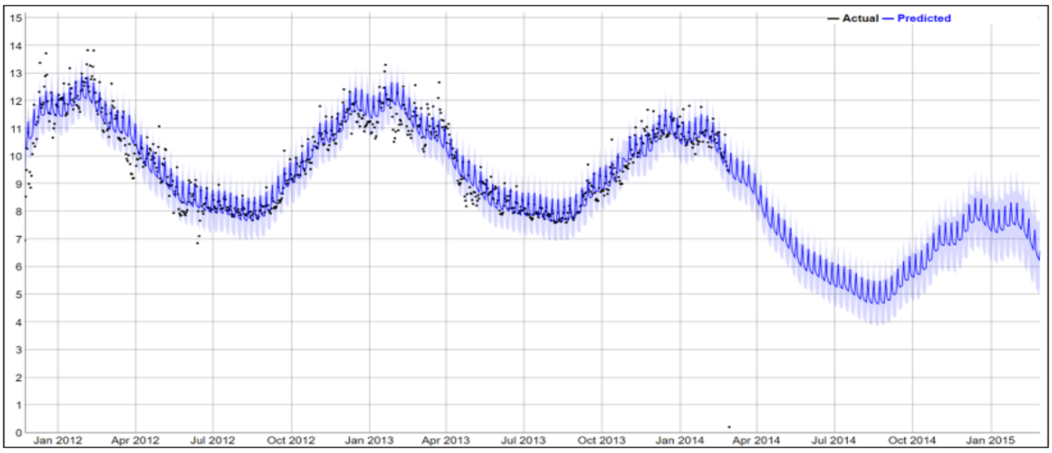

Figure 8: Real v/s Forecast Outcome Prophet Model

Evaluation - The PROPHET model performance was assessed by the Y and Yhat values, where Y stand for the real electricity load consumption values and the Yhat stands for the forecasting values. It was observed 2.89 as the RSME value and 8.35 as the MSE, furthermore, the 2.42 is the MAE score found and the 4.27 is the MAPE score.

RMSE	MSE	MAE	MAPE
2.89	8.35	2.42	4.27

Table 4: Evaluation Metrics Result of PROPHET Model

5.4.4 Experiment 4 - Long Short-Term Memory (LSTM)

The pre-processed and transformed data has been retrieved from R Studio as a CSV file, this CSV file has been imported into the python programming environment for the implementation of the LSTM model.

The scaling of data is an important aspect considered before the LSTM modelling. Therefore, before breaking the complete data into a train and test set, scaling of data has been carried out. Scaling of data is required for maintaining the variance in attributes even in the same range. In another aspect, if one attributes has greater variation in comparison to other attributes, it can impact the performance of the forecasting model. This study considers a univariate data-set consisting of date as index and regular electricity consumption as the forecasting attribute. Therefore, scaling of the electricity load consumption attribute has been done using MinMaxScaler() function of the python programming to reduce the variation of data within the defined range and determine the accuracy of the model performance. Next, the scaled data was split in a ratio of 70:30 into a train and test set. The train set has been considered for the LSTM model training and test set has been used for the evaluation of the forecasting outcomes.

In an aim to minimize the complexities, the following conditions have been used for a simple yet robust network. The implemented model is composed with one input layer, one hidden layer, and one output layer with a linear activation function. The model works upon 4 LSTM neuron, Mean square error as the loss function along with 'Adam' optimiser. Firstly the validation split of 0.3 was done which has split the training samples in Train on 403 samples, validate on 173 samples format. Next, The batch size considered was 15 and the epochs considered was 50. As a whole the neurons compute the weighted sum of inputs, add bias and pass value to the activation function that generates output. Finally, it was observed that during the model training at the final epoch, loss: 0.0014 - val_loss: 0.0012.

Once the model training is completed, it is than used to the generate forecast values using model.predict() function. Restoring of the scaled data to their original format is one of the key steps, that has been carried out after the forecast value is generating. As, this help in computation of the evaluation matrix.

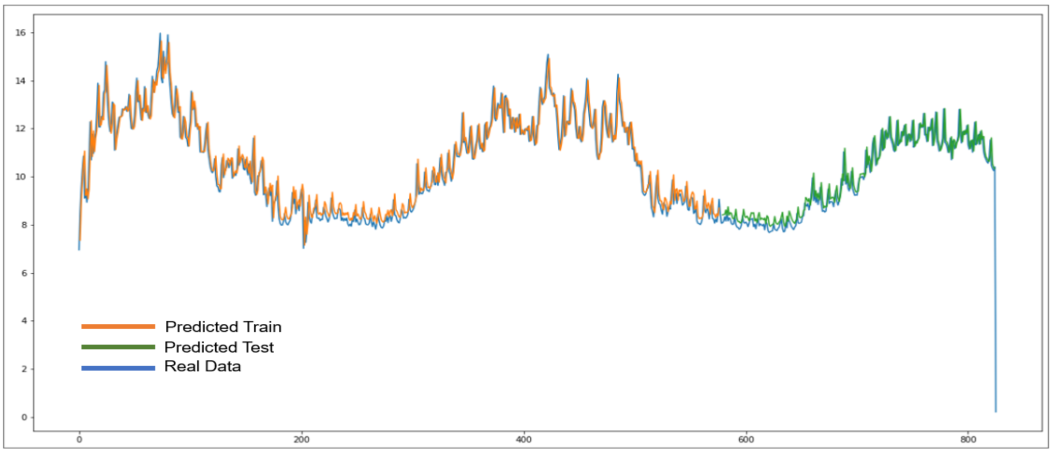

Figure 9: Real v/s Forecast Outcome LSTM Model

The graph present in Figure 9 represent the real value in comparison to the forecasting outcome of LSTM model, blue line indicates the real value while the orange and the green line over the blue line indicate the forecasting outcome. It is being observed that orange and green line quite closely match the real data.

Evaluation - The sklearn.metrics package of the python programming is being used to evaluate the performance of the LSTM model. The performance was evaluate by the 30% of data that was masked was a test data against the forecasting outcome. It noticed that the RSME value is 0.40 and the MSE value is 0.16, likewise the MAE is 0.29 and the MAPE is 3.06.

RMSE	MSE	MAE	MAPE
0.40	0.16	0.29	3.06

Table 5: Evaluation Matrics Result of LSTM Model

6 Model Performance Comparison

The analysis has been carried out using four separate models, it is therefore necessary to evaluate the models outcome with appropriate measures in order to determine the best fit model. A standard procedure of splitting the complete data into 70:30 ratio for model training and testing has been followed, all throughout the study. Ultimately, the computations were carried out by comparing the real value against the predicted results. The measure considered to evaluate the model performance in this study are Root Mean Square Error (RSME), Mean Square Error (MSE), Mean Absolute Error and Mean Absolute Percentage Error (MAPE). While many conflicts of opinion have been rendered about selecting a technique over a specific approach to access the model performance. However, a study by (Chai and Draxler; 2014) made it evident that the use of RSME and MSE techniques are effective in quantifying the machine learning models performance, also relying upon a single major for the assessment of machine learning models can sometime fail to detect the model performance flaws.

Models	RSME	MSE	MAPE	MAE
ARIMAX	0.76	0.57	0.47	6.21
SARIMAX	0.77	0.60	0.50	6.50
PROPHET	2.89	8.35	2.42	4.27
LSTM	0.40	0.16	0.29	3.06

Table 6: Models Performance Measurement

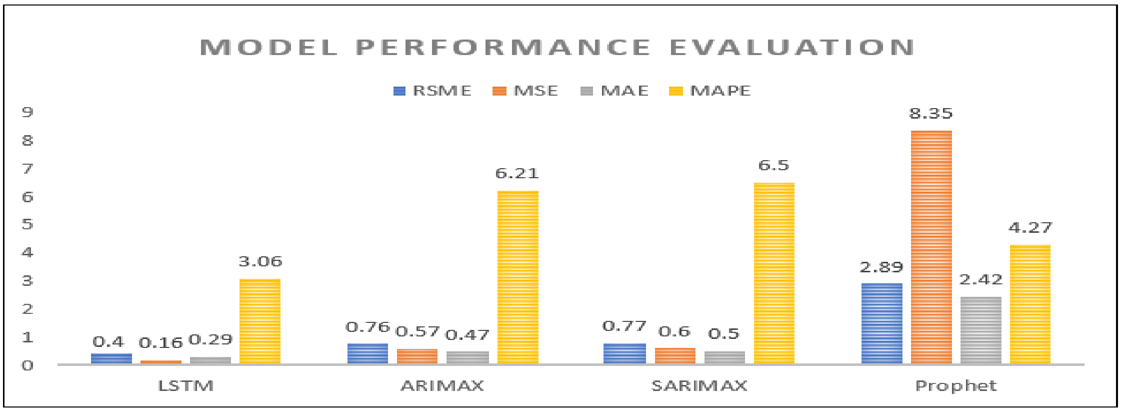

Figure 10: Model Evaluation

By comparing the accuracy measures as demonstrated in Table 2, it can be seen that the RSME score of the prophet model is maximum at 2.89 and MAPE of 4.27%, whereas the RSME scores of ARIMAX and SARIMAX is quite similar to each other at 0.76 and 0.77. The ARIMAX and SARIMAX models however presented highest MAPE of 6.21% and 6.50% respectively. Finally, it can be determined that the LSTM is the best-performing model with the lowest RSME score of 0.40 and the Lowest MAPE of 3.06%. Moreover, in terms of model accuracy by these error measures it was concluded that the performance of the LSTM model has been satisfactory.

7 Discussion

This research in specific offers insights into the implementation of different machine learning and a deep learning models in forecasting residential consumer electricity load demand. Electricity Load Forecasting enables Electricity transmission and distribution companies in taking strategic decisions including acquisition and generation of electrical energy, load switch-over as well as development of infrastructure. The project was carried in different phases such as,

1. Choosing a compatible data set. The data-set for this analysis was curated from a data repository of Kaggle data repository, consisting of well distributed attributes with minimum missing values of London residential house energy load consumption smart meter readings and weather reports.

2. Next, in the pre-processing and the data transformation phase, the missing values were initially removed from the data set and average electricity usage of the entire residential buildings was computed by grouping data according to the date. Additionally, further computation was carried out to determine the average daily temperature.

3. Finally, in the model implementation phase there were a few complications faced such as, data for ARIMAX and SARIMAX had to include seasonally optimized data for an enhanced performance of these models. Therefore, this has been done by utilizing various packages present in R Programming Language. Similarly, scaling of data was performed in python programming prior to LSTM model implementation and reverse scaling of the forecasting output was done to compute the model performance. Also, in case of the Facebook PROPHET model prerequisite the naming convention of forecasting attributes has been modified to 'y' and the date attributes to 'ds' accordingly.

8 Conclusion and Future Work

The ultimate goal of this analysis is to forecast the electricity load requirements of residential customers at London through different statistical techniques, by utilizing evidence of past energy usage of the customers and daily average temperature. The model considered for the implementation were ARIMAX, SARIMAX, PROPHET and LSTM. Although the four models have different configurations, it is witnessed that a well processed dataset, suitably designed framework, and a well-implemented procedure have resulted to favourable performance that is available from all the four models. The performance assessment of the models by root mean square error and mean absolute performance error computation clearly indicated that PROPHET is the least performer out of all four models with high error score. However, the performance of ARIMAX and SARIMAX model were quite similar to each other and satisfactory. Ultimately it is found that LSTM model outperformed all the other models as it rendered a lowest RSM and MAPE score.

In future, implementation of multivariate LSTM model can therefore be taken into consideration with several other features such as household energy ratings. Additionally, the forecasting scope can also be extended to cover commercial and industrial buildings electricity consumption.

9 Acknowledgement

In the course of developing this research project, I would also like to thank my Research Supervisor Dr. Manaz Kaleel for his extremely valuable support and feedback. I would also like to thank my family and friends for their continuous encouragement.

References

Afshar, O., Saidur, R., Hasanuzzaman, M. and Jameel, M. (2012). A review of thermodynamics and heat transfer in solar refrigeration system, *Renewable and Sustainable Energy Reviews* **16**(8): 5639–5648.

Almeshaiei, E. and Soltan, H. (2011). A methodology for electric power load forecasting, *Alexandria Engineering Journal* **50**(2): 137–144.

Asha, J., Rishidas, S., SanthoshKumar, S. and Reena, P. (2019). Analysis of temperature prediction using random forest and facebook prophet algorithms, *International Conference on Innovative Data Communication Technologies and Application*, Springer, pp. 432–439.

Bouktif, S., Fiaz, A., Ouni, A. and Serhani, M. A. (2018). Optimal deep learning lstm model for electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches, *Energies* **11**(7): 1636.

Chai, T. and Draxler, R. R. (2014). Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature, *Geoscientific model development* **7**(3): 1247–1250.

Chinnathambi, R. A., Plathottam, S. J., Hossen, T., Nair, A. S. and Ranganathan, P. (2018). Deep neural networks (dnn) for day-ahead electricity price markets, *2018 IEEE Electrical Power and Energy Conference (EPEC)*, pp. 1–6.

Drysdale, B., Wu, J. and Jenkins, N. (2015). Flexible demand in the gb domestic electricity sector in 2030, *Applied Energy* **139**: 281–290.

Elamin, N. and Fukushige, M. (2018). Modeling and forecasting hourly electricity demand by sarimax with interactions, *Energy* **165**: 257–268.

Guo, Z., Zhou, K., Zhang, X. and Yang, S. (2018). A deep learning model for short-term power load and probability density forecasting, *Energy* **160**: 1186–1200.

Hayn, M., Bertsch, V. and Fichtner, W. (2014). Electricity load profiles in europe: The importance of household segmentation, *Energy Research & Social Science* **3**: 30–45.

Le, T., Vo, M. T., Vo, B., Hwang, E., Rho, S. and Baik, S. W. (2019). Improving electric energy consumption prediction using cnn and bi-lstm, *Applied Sciences* **9**(20): 4237.

Mariscal, G., Marban, O. and Fernandez, C. (2010). A survey of data mining and knowledge discovery process models and methodologies, *The Knowledge Engineering Review* **25**(2): 137.

Meier, H. and Rehdanz, K. (2010). Determinants of residential space heating expenditures in great britain, *Energy Economics* **32**(5): 949–959.

Mocanu, E., Nguyen, P. H., Gibescu, M. and Kling, W. L. (2014). Comparison of machine learning methods for estimating energy consumption in buildings, *2014 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)* .

Nateghi, R. and Mukherjee, S. (2017). A multi-paradigm framework to assess the impacts of climate change on end-use energy demand, *PloS One* **12**(11): e0188033.

Pirbazari, A. M., Farmanbar, M., Chakravorty, A. and Rong, C. (2020). Short-term load forecasting using smart meter data: A generalization analysis, *Processes* **8**(4): 484.

Pérez-Lombard, L., Ortiz, J. and Pout, C. (2008). A review on buildings energy consumption information, *Energy and Buildings* **40**(3): 394–398.

Ramírez-Mendiola, J. L., Grünewald, P. and Eyre, N. (2017). The diversity of residential electricity demand–a comparative analysis of metered and simulated data, *Energy and Buildings* **151**: 121–131.

Santika, G. D., Mahmudy, W. F. and Naba, A. (2017). Electrical load forecasting using adaptive neuro-fuzzy inference system, *Int. J. Adv. Soft Comput. Appl.* **9**(1): 50–69.

Siami-Namini, S., Tavakoli, N. and Siami Namin, A. (2018). A comparison of arima and lstm in forecasting time series, *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 1394–1401.

Staffell, I. (2017). Measuring the progress and impacts of decarbonising british electricity, *Energy Policy* **102**: 463–475.

Tarsitano, A. and Amerise, I. L. (2017). Short-term load forecasting using a two-stage sarimax model, *Energy* **133**: 108–114.

Ulyah, S., Mardianto, M. et al. (2019). Comparing the performance of seasonal arimax model and nonparametric regression model in predicting claim reserve of education insurance, *Journal of Physics: Conference Series*, Vol. 1397, IOP Publishing, p. 012074.

Ushakova, A. and Murcio, R. (2018). Interpreting smart meter data of uk domestic energy consumers, *Consumer Data Research* p. 120–137.

Wang, C., Bäck, T., Hoos, H. H., Baratchi, M., Limmer, S. and Olhofer, M. (2019). Automated machine learning for short-term electric load forecasting, *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, IEEE, pp. 314–321.

Wang, W.-C., Chau, K.-W., Cheng, C.-T. and Qiu, L. (2009). A comparison of performance of several artificial intelligence methods for forecasting monthly discharge time series, *Journal of hydrology* **374**(3-4): 294–306.

Wee, C. K. and Nayak, R. (2019). Adaptive load forecasting using reinforcement learning with database technology, *Journal of Information and Telecommunication* **3**(3): 381–399.

Weron, R. (2014). Electricity price forecasting: A review of the state-of-the-art with a look into the future, *International Journal of Forecasting* **30**(4): 1030–1081.

Xydas, S., Marmaras, C., Cipcigan, L. M., Hassan, A. and Jenkins, N. (2013). Electric vehicle load forecasting using data mining methods.

Zhao, H.-X. and Magoulès, F. (2012). A review on the prediction of building energy consumption, *Renewable and Sustainable Energy Reviews* **16**(6): 3586–3592.

Zunic, E., Korjenic, K., Hodzic, K. and Donko, D. (2020). Application of facebook's prophet algorithm for successful sales forecasting based on real-world data, *arXiv preprint arXiv:2005.07575* .