JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

ELECTRICITY CONSUMPTION PREDICTION USING MACHINE LEARNING MODELS

[1]K. Kushwanth Sai, [1]B. Prasanthi, [1]K. Anil Kumar, [1]K. Priyanka, [2]Dr P. Padmaja

[1]Student, [2]Professor
[1]Department of Information Technology,
[1]Anil Neerukonda Institute of Technology and Sciences, Sangivalasa, Visakhapatnam, Andhra Pradesh, India.

Abstract: Reliable power demand predictions that account for the effects of extreme weather events is necessary for guiding electricity supply operation and utility resource planning and improving energy security and grid resilience. Three typical data-driven models are used to forecast city-scale daily power demand: linear regression models, machine learning models for time series data, and machine learning models for tabular data. Seven models are applied to city-scale power demand data from three Indian metropolises: Hyderabad, Mumbai, and Delhi. The results demonstrate that seven models can forecast daily power usage in the metropolitan region with a coefficient of variation of the root mean square error (CVRMSE) of less than 10%. The lightGBM produces the most accurate findings, with a CVRMSE of 6.5% for Mumbai, 4.6% for Hyderabad, and 4.1% for the Delhi metropolitan region on the test dataset. According to the findings, weather-sensitive components account for 30%-50% of daily power demand. Every degree Celsius rise in summer ambient temperature results in approximately 5% (4.7% in Mumbai, 6.2% in Hyderabad, and 5.1% in Delhi) greater daily power consumption than the base load in the three metropolitan cities.

Index Terms **– City-scale electricity usage, Decomposed time-series modelling, Gradient Boosting Trees, temperature-sensitive energy demand, Machine Learning prediction.**

I. INTRODUCTION

Predicting city-scale electricity demand can help with power-generating resource planning, energy conservation programme assessment, monitoring greenhouse gas emissions, system infrastructure, and reserve requirements analysis. Therefore, understanding building energy usage at the city scale is crucial for developing worldwide urban sustainability, carbon reduction, and power efficiency. Electricity use in cities is temperature sensitive. Since cooling and heating buildings are substantial energy uses in cities, largely reliant on external air temperature, room temperature, alongside variables like population and income, is a critical driver of city-scale energy consumption. Examining temperature-sensitive city-scale power usage is becoming critical as climate change causes more frequent, intense, and prolonged severe weather conditions such as heat waves. To produce successful solutions, academics, energy planners, and politicians must prioritise climate change adaptation. Similarly, these stakeholders must comprehend how power production and transmission infrastructure may be better equipped for high-demand events to improve energy security and resilience in the face of climate change.

Hou et al. (2014) researched how rising ambient temperatures affected power use in Shanghai. They contended that if the existing power consumption trend does not change, the anticipated temperature increase indicates an increase in summer power demand and a decrease in winter demand. Similarly, it was anticipated that global warming and related peak demand increases in California would entail up to 38% more peak generating capacity and up to 31% more transmission capacity by 2099. A heat wave in California in August 2020 caused a power supply shortfall due to increased air-conditioning consumption, and California residents endured rotating power outages. As policymakers considered this disruptive occurrence, the first action specified in the Answer Letter was to revise the power demand projection for climate change, including severe weather conditions and their related load implications.

Other variables, such as unforeseen public health incidents, might impact city-scale power demand and the ambient meteorological condition. According to research performed in Brazil, the COVID-19 epidemic altered Brazilian power consumption habits. Consequently, according to the regional economic structure, power consumption in Brazil was lowered by 7%-20%, with the industry-dominated area being less affected. Another European study discovered that the severity and duration of lockdown measures increased societal electricity use. Power usage profiles during the pandemic in countries with stringent regulations, such as Spain and Italy, are like pre-pandemic weekend accounts for the same time frame in 2019. In countries with fewer new restrictions, such as Sweden, the reduction in electricity consumption was lower. Electricity usage may be utilised to track the shutdown's economic consequences in real time. The overall power usage decline in Switzerland was determined to be 4.6%. For instance, the Canton of Ticino saw a 14.3% decline after stricter restrictions were added on top of federal laws. Top-down and bottom-up methodologies are two broad ways to model city-level electricity use. Top-down models look at city-level energy usage on a macro scale, neglecting the specifics of end applications. The top-down technique is commonly used to determine the links between energy use and socioeconomic and demographic parameters. The bottom-up technique creates simulation models for individual units, which are then aggregated to

compute macro-level energy use. The simulation unit does not have to be a single end user, as there may be millions of users in the city. Instead, it could be a cluster/group of end consumers with similar traits or tendencies.

Because developing a physics-based model at the macro level is difficult (for example, because establishing a physics-based model involves a significant quantity of input data with considerable uncertainty), top-down modelling typically employs data-driven methodologies. In contrast, the bottom-up method allows for the development of either a physics-based or a data-driven model. Wang et al. (2015) created a bottom-up physics-based model to estimate China's state-level heating energy consumption [15]. Kontokosta and Tull (2017) created a bottom-up data-driven algorithm to forecast city-scale building energy usage in New York [16]. The bottom-up approach allows city-scale building retrofit analysis covering individual buildings with varying characteristics and baseline performance. This approach cannot be conducted using top-down statistical-based methods [17]. It is difficult to say whether the top-down or bottom-up approaches are better, as both methods have strengths and weaknesses [18]. Which approach is more suitable depends on the application and objective of the modelling (i.e., fit-for-purpose).

Project Scope and Direction:

The forecasting of city-scale electricity has various applications, and several research papers have been published on this topic. However, upon analysing existing studies, a significant research gap was identified. Despite the numerous methodologies proposed, none of the research discloses open-source data, models, or code for public review or reuse. Consequently, these methods limit the potential for summary and comparison, particularly when comparing traditional linear regression approaches to developing machine learning algorithms.

This research paper utilizes a top-down data-driven approach to predict daily power use at the city level. Three techniques for predicting daily energy consumption at the city scale were investigated, including linear regression models, machine learning models for time series data, and machine learning models for non-time series data.

Impact, Significance and Contribution:

The Electricity consumption prediction have positive impacts and benefits in the power sector:

1. Developing productivity
2. Protection against electricity losses and decreased production
3. Control of electricity usage
4. Prediction of usage in future

II. RELATED RESEARCH

Corgnatiet al. (2013) suggested a data-driven method with well-defined input and output attributes. Based on this information, system parameters would be assessed, and a mathematical model would be created. Numerous previous studies on this data-driven machine-learning technology have been carried out.

Fu et al. [2015] proposed using a Support Vector Machine, ML approach to estimate load at a building's air conditioning, lighting, electricity, and other systems based on weather predictions and hourly electrical demand input. Using the SVM technique, the total electrical demand was projected with an RMSE of 15.2% and a mean bias error (MBE) of 7.7%.

Valgaev et al. [2016] provided a power consumption estimate for a smart building using the k-Nearest Neighbor model. To develop the k-NN forecasting method, historical data and their successors were accumulated. Because it only recognises similar occurrences in an ample feature space, the k-NN technique has limitations in predicting future value. As a result, it must be supplemented during the workday by temporal information identification, with predictions made for the next 24 hours.

El Khantach et al. [2019] forecasted short-term load using five machine-learning techniques. He ultimately developed a 24-time series for every hour after decomposing historical data into time series for every hour of the day. The five machine learning methods are the multi-layer perceptron, support vector machine, radial basis function regressor, REPTree, and Gaussian process. The investigation was carried out utilising data from Moroccan electrical load statistics. The MLP technique was judged to be the most accurate, with a MAPE percentage of 0.96, followed by SVM, which, although not as precise as MLP, was relatively better than the others.

According to Gonzalez-Brione et al. [2019], while energy consumption prediction is often performed using a classification-based machine learning methodology, the regression method may also be applied. The data was studied using Linear Regression, Support Vector Regression, Random Forest, Decision Tree, and k-Nearest Neighbor to develop a prediction model. One day before, the research criteria included electricity use as a separate feature. The LR and SVR models topped the others, with an accuracy of 85.7%.

Newgard and Lewis [2015] provide many imputation procedures, including Mean Value Imputation, Last Observation Carried Forward, Maximum Likelihood Estimate, and Multiple Imputation. The mean value imputation replaces missing data with the dataset's mean value. This method, however, is unsuitable for non-strictly random data since it introduces inequality into the data. Multiple Imputation and Maximum Likelihood Estimation were the most difficult methodologies shown. With each repeat, the Multiple Imputation approaches gradually replace missing data. This approach employs statistical analysis based on visible data to deal with the uncertainty the missing component introduces. Multiple Imputation Using Chained Equations is a well-known MI method. The Maximum Likelihood Estimate achieves replacement by the assumption by first establishing the parameters and boundaries based on the data distribution. This imputation approach was employed in Probabilistic Principal Component Analysis.

III. METHODOLOGY

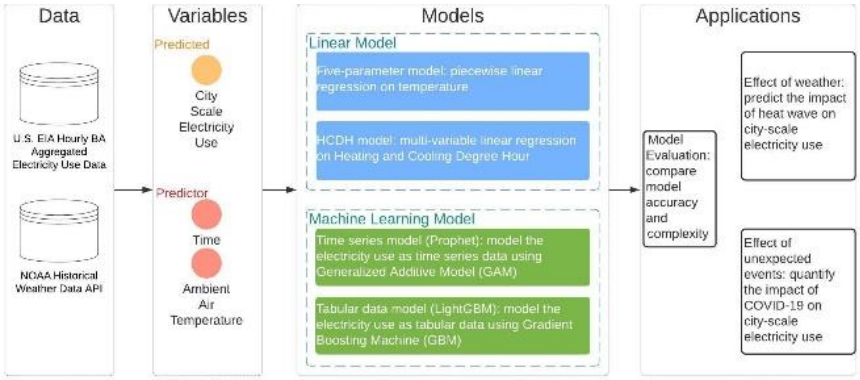


Fig1 Electricity consumption prediction Methodology

In this section, the existing city-level electricity usage models are reviewed, followed by the introduction of the models developed in this study. The workflow of the study can be seen in Figure 1. Seven data-driven models were developed to predict daily electricity use at the city scale, belonging to three different modelling approaches: linear models, machine learning models for time-series data, and machine learning models for tabular data. To begin with, two linear models were developed due to their simplicity and interoperability: ASHRAE's Five-Parameter piecewise linear regression model and the Heating Cooling Degree Hour model. As for the machine learning models, three conventional models were developed as baselines: Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (NN).

In addition, two advanced machine learning models, Generalized Additive Model (GAM) and Gradient Boosting Machine (GBM), were developed and compared. GAM and GBM model the electricity consumption data differently, with GAM modelling electricity usage as time series data and GBM modelling electricity usage as tabular data. The main difference between time series modelling and tabular data modelling is how the temporal information is encoded. Given the clear weekly and yearly cycles in electricity usage, it is natural to use time series modelling to predict electricity usage. The time series model treats electricity use as time series data, representing temporal information using an evolving index.

Electricity usage data can also be represented in a tabular format, requiring the inclusion of extra features like the day of the week and month of the year to capture weekly and yearly cycles. The tabular model uses additional features to represent temporal information, such as the hour and day of the year. In time series models, the data sequence contains temporal information and cannot be altered. This enables the shuffling of the data sequence in tabular models. The conventional machine learning models (RF, SVM, NN) use the tabular data format to model electricity usage.

This study aims to predict city-scale daily electricity use, including energy consumption for buildings, transportation, industry, and public services in the city and nearby rural areas. The ambient air temperature plays a significant role in determining electricity usage. We compare the time series and tabular data models with linear and other baseline machine learning models.

Linear models:

Linear models are commonly utilized in city-level energy modeling due to their simplicity and interoperability. These models utilize a linear relationship to regress the observation and independent variables. Their effectiveness has been demonstrated in various studies such as Lindsey et al. (2011) who developed a linear model to predict the city-level transportation energy usage and greenhouse gas emissions in Chicago. Additionally, Kuusela et al. (2015) developed a multi-variable linear regression model to predict the energy consumption at a neighborhood scale [20]. Linear models have even been shown to perform well in predicting the energy output of complex systems such as large-scale ground source heat pumps [21].

One of the main reasons why linear models are widely used is because of their interoperability, as the coefficients can be used to validate and explain the models. Although linear models may not capture non-linear relationships that are common in the real world, their simplicity and understandability make them a suitable choice as a baseline model for benchmarking against more complicated non-linear models. In this study, two linear models were developed to predict temperature-sensitive electricity usage on a city scale. The first linear model used in the study is ASHRAE's change-point model, which was proposed by ASHRAE in the 1990s. The change point model uses five parameters ($\beta base$, Th, Tc, βh, βc) to characterize the relationship between energy usage and ambient temperature. The base load, represented by $\beta base$, is the lowest energy usage when the outdoor temperature is within the range of [Th, Tc]. When the outdoor temperature falls below the heating change point Th, city-level energy usage increases in response to the temperature decrease due to the increase in heating demand. Figure 2 and Equation 1 depict the change point model.

On the other hand, when the temperature outside surpasses the cooling threshold Tc, there is a surge in the demand for cooling, leading to an increase in energy usage at the city level. The city-level load sensitivity to temperature change is characterized by the slopes on the cooling (βc) and heating (βh) sides. As the focus is solely on electricity consumption, the value of βh would be lower than βc, because numerous buildings utilize natural gas for heating, while most air-conditioned buildings use electricity for cooling. The

ASHRAE's five-parameter change point model is extensively employed to estimate building-level energy consumption, benchmark building energy performance, and, in this study, we used the five-parameter change point (5-p) model to predict city-level energy usage.

$$load(T) = \beta base + \beta h \times (Th - T), if\ T < Th$$

$$\beta base \quad , if\ Th < T < TC$$

$$\beta base + \beta c \times (T - Tc), if\ TC < T \quad \text{(Equation 1)}$$

The second linear model used in this study is the Heating/Cooling Degree Hour (HCDH) model. This method is widely used in the heating, ventilating, and air conditioning (HVAC) industry to estimate heating and cooling energy requirements [27]. The heating and cooling degree day is a significant proxy variable used to quantify the impact of climate change on electricity demand [28]. The HDH and CDH are calculated as the cumulative sum of the difference between the ambient temperature and the heating base load temperature (Tbh) and cooling base load temperature (Tbc) shown in Equation 2. When the outdoor temperature drops below Tbh, heating is triggered, and the cumulative sum of the difference between the outdoor and base temperatures ($Tbh - Ti$) is a good indicator of the required heating. HDH and CDH are widely used for building energy demand estimation [29], determining the building's thermal insulation [30], and other purposes. In this study, the daily city-level energy usage was regressed as a linear function of the HDH and CDH.

$$HDH = \sum \max(0, (Tbh - Ti))\ 24\ i{=}1$$

$$CDH = \sum \max(0, (Ti - Tbc))\ 24\ i{=}1$$

$$load(T) = \beta 0 + \beta 1 \times HDH + \beta 2 \times CDH \quad \text{(Equation 2)}$$

A common challenge encountered in developing both the five-parameter and HCDH models involves the careful selection of the change temperature (Th, Tc in the five-parameter model) and the base temperature (Tbh, Tbc in the HCDH model), as reported in [31]. This study's approach for selecting these temperatures involved identifying which sets of change or base temperature could yield the most precise linear model. To achieve this, the best change temperatures for the five-parameter model were determined using the scipy.optimize.curve_fit function [32], while the best base temperatures for the HCDH model were chosen using a brute force search.

Machine learning model for time-series data:

The Autoregressive Integrated Moving Average (ARIMA) is a time-series modelling technique that is known as the oldest. ARIMA predicts a time-series variable yt using its own lagged values ($yt{-}1$, $yt{-}2$…) and previous prediction error ($\varepsilon t{-}1$, $\varepsilon t{-}2$, …). The prediction error is the difference between the predicted value $\hat{y}i$ and the true value yi. ARIMA has been extensively used to predict energy demand, such as natural gas demand in Turkey and electricity demand in Lebanon. Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) have been used to decide the order of the ARIMA model and validate electricity demand load forecasting models. ARIMA has also been combined with other techniques, such as wavelet transform, to improve its prediction accuracy. Recurrent Neural Network (RNN) and its variant Short-Term Long Memory (LSTM) are other mainstream machine learning models for time-series data. These models use a different approach to encode the time dependency and input a state from the previous time step to predict yt. Recent advancements in deep learning algorithms and computational power have led to the widespread use of neural network-based approaches to predict energy usage. More than 40 papers have used a neural network-based approach for short, medium, and long-term load forecasts. Recent publications have used RNN and LSTM to predict electricity consumption for commercial and residential buildings and district-level energy demand. Studies comparing ARIMA and RNN/LSTM have found that LSTM outperforms ARIMA in building load forecast because it can capture non-linear relationships between time-series data and exogenous variables.

This paper discusses a novel approach for time-series modeling to predict electricity usage at the city level. The study employs two techniques, namely ARIMA and RNN. The modeling is performed using a decomposed time-series model comprising three main components: exogenous variables, trend, and seasonality. Equation 3 shows that the linear function of heating and cooling degree hours ($f(Tempt)$) is one of the components. The trend function (g(t)) models non-periodic changes in the time series, and s(t) models yearly and weekly seasonality changes. To implement this model, Facebook's Prophet, which is an open-source software, was utilized. For additional details on the implementation, refer to the paper [43].

$$loadt = f\ (Tempt) + g(t) + s(t) + \varepsilon t \quad \text{(Equation 3)}$$

The decomposed time-series model was chosen for two reasons. Firstly, it has a proven track record in predicting temperature, financial markets, and daily COVID-19 cases in Bangladesh. Secondly, the model can separate the effects of various factors (such as temperature-dependent load and seasonal time-dependent periodical load) and allow for the observation of the isolated impact of unexpected public health events on city-level demand. This is the first time the model has been applied to predict city-level energy consumption, as far as the authors are aware. The decomposed time series model has the advantage of breaking down time series data into different components, each with distinct implications. The temperature-sensitive load f (Tempt) is often linked to HVAC use, which varies with temperature. The periodic load s(t) captures load variations as a function of time, such as the effect of holidays on city-level electricity usage. The non-periodic load g(t) represents the remaining load variation, which may be due to short-term factors (such as the COVID-19 pandemic) or long-term trends (such as improving building thermal properties and equipment energy efficiency). The decomposition results can provide insights into the magnitude of each component and identify the primary driving factors.

It is important to note that in developing the time series decomposed model, the daily heating and cooling degree hour was used as the intermediate variable instead of the daily mean temperature. This is because the f (Tempt) term in the decomposed model is a linear function, and therefore, a monotonous one. The relationship between city-wide daily electricity use and ambient mean temperature follows a U-shaped curve, where high electricity use occurs when the temperature is either very low or very high. Since a monotonous function cannot capture this U-shape relationship, the daily heating and cooling degree per hour is used as the regressor in the model, as the relationship between electricity use and heating and cooling degree per hour is monotonous.

Machine learning model for tabular data (non-time-series data):

City-level electricity consumption can also be modelled as tabular data. New features must be added to capture energy usage's timing and periodic behaviour. For instance, to encode the weekly and yearly cycles, two new features—the day of the week and the month of the year—must be added as input variables.

Numerous studies utilize tabular data to model time series energy usage. Machine learning algorithms for this type of data can be categorized into three major groups: neural network-based, decision tree-based, and other methods. Neural network-based approaches, also known as artificial neural networks, feedforward neural networks, or multi-layer perception, use neural networks to tackle regression or classification tasks. These approaches have the advantages of being easily parallelizable and applicable to various tasks. For instance, Fernández et al. (2011) utilized an NN-based approach to predict building load. Decision tree-based approaches consist of Classification and Regression Tree, Random Forest, and Gradient Boosting Machine, which employ ensemble learning techniques by combining multiple decision trees to enhance model accuracy and robustness. Tso and Yau (2007) applied CART to forecast building energy usage in Hong Kong, while Roth et al. (2019) developed RF and GBM models to anticipate building energy consumption in New York City. Additionally, other tabular data algorithms, such as Support Vector Machine, k-Nearest Neighbors, and k-means, have also been used to model energy usage. For example, Li et al. (2017) applied the Support Vector Machine approach to forecast community-level renewable generation, while Al-Qahtani and Crone (2013) used k-Nearest Neighbors to predict electricity demand in the United Kingdom. Moreover, Fonseca and Schlueter (2015) utilized k-means to predict district-level building energy usage in Zurich. Finally, Kontokosta and Tull (2017) discovered that the best machine learning algorithm for energy usage prediction could depend on the geographical resolution, with SVM performing the most accurately at the building level and Linear regression outperforming other methods at the zip code level.

Others Machine Learning Models:

This research paper considers four tabular data modelling algorithms, namely RF, SVM, NN, and GBM. The first three algorithms, RF, SVM, and NN, are used as baseline algorithms while GBM is discussed in detail. GBM has gained significant attention in recent years due to its superior performance in various machine learning competitions and studies. For instance, in the ASHRAE Great Energy Predictor III competition, all the top six teams used GBM in their final predictors. Moreover, previous studies have demonstrated that GBM outperforms several other popular machine learning algorithms, such as Ridge regression, Lasso regression, Elastic Net, Support Vector Machine, Random Forest, vanilla Deep Neural Network, and Long Short-Term Memory, in predicting building loads. Thus, GBM is chosen as the state-of-the-art algorithm and serves as a representative of the tabular data modelling approach in this study.

The initial stage involves utilizing the time index for producing intermediary variables that encode temporal data. Numerous GBM implementation packages are available, and Microsoft's lightGBM [55] was employed in this case due to its user-friendly interface and comprehensive documentation. Proper hyper-parameter tuning is essential to prevent over-fitting or under-fitting while training a GBM.

IV. RESULTS

Model accuracy is considered one of the most important criteria for comparing different data-driven models. To compare the models accurately, the model was trained using data from July 2015 to June 2018, and the model's performance was evaluated using data from July 2018 to June 2019. The data from 2020 was excluded to avoid any unexpected events affecting the comparison. Seven data-driven models were developed, including two linear models (five-parameter change point model, Heating and Cooling Degree Hour model), three conventional machine learning models (Random Forest, Support Vector Machine, Artificial Neural Network), a time-series decomposed model, and Gradient Boosting Machine model. The models were trained on the training set and then evaluated using three different metrics: mean absolute error (MAE), root mean squared error (RMSE), and cross-validation root mean squared error (CVRMSE). Additionally, Figure 1 displays the CVRMSE of the seven algorithms across three metropolitan areas.

Top-down data-driven models can accurately predict the electricity demand at a city level. All seven of these models can predict the daily electricity usage of a city with over 90% accuracy. While overfitting is acceptable for Hyderabad or Delhi, in Mumbai, the model's performance on the test dataset deteriorated significantly more than the other two regions. The models performed similarly well on both the train and test datasets in the other two regions. Two possible reasons for this are that the electricity usage behavior of Mumbai changed in the test dataset or that the data-driven model identified and included other hidden factors that significantly influence Mumbai's electricity demand.

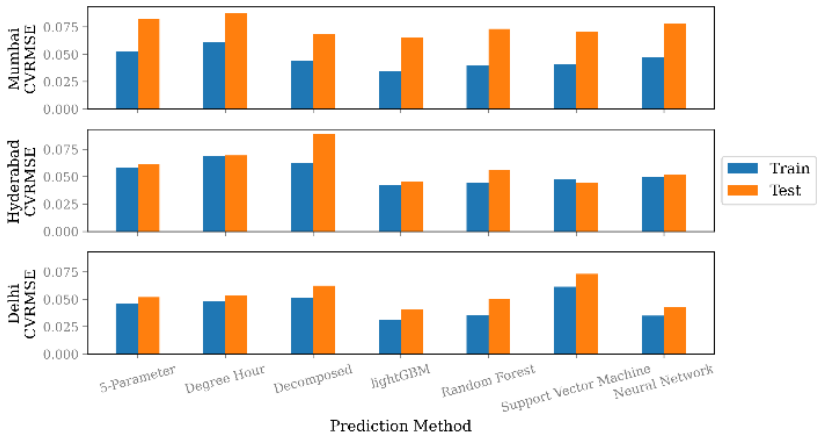
Simple linear regression models, such as piecewise and multivariate models, can provide satisfactory predictions. The five-parameter change point model outperformed the Heating and Cooling Degree Hour model in all three metropolitan areas. The CVRMSE of the five-parameter model ranged from 5.2% to 8.2%, while the CVRMSE of the HCDH model ranged from 5.4% to 8.8%. Moreover, the five-parameter change point model is more straightforward to implement because it does not require determining the best-performing heating and cooling base temperatures, which may vary among cities with different weather and electricity use behaviour.

Machine Learning models such as RF, SVM, and NN exhibit comparable accuracy levels. However, the effectiveness of these models is influenced by the dataset they are applied to, even when utilizing the same hyper-parameters and model architecture. For example, while SVM may yield superior results in Mumbai and Hyderabad, it may perform poorly in Delhi. Therefore, the ability of the model to generalize is brought into question.

The GBM model demonstrates superior performance compared to other baseline machine learning models (RF, SVM, and NN) and the decomposed model in all three metropolitan areas. The decrease in CVRMSE ranges from 0.4% (Mumbai) to 4.3% (Hyderabad). In this study, the careful encoding of temporal information using two variables (month of year and day of the week) suggests that electricity usage might not necessarily be modeled as time-series data. Time-series modeling employs the sequential ordering of input data to represent temporal information and is less robust to missing data compared to tabular data models. Imputation of missing data in time-series modeling adds complexity to data preprocessing. However, tabular data models utilize additional features (e.g., day of the week) to encode temporal information and are not as affected by missing data.

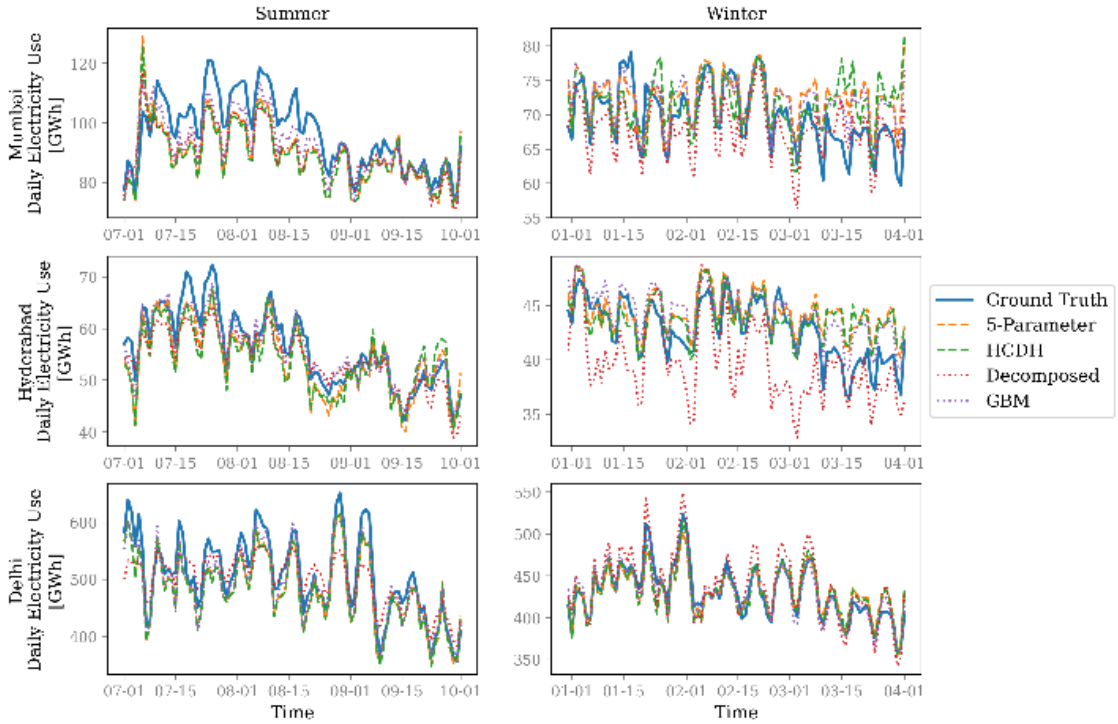
Comparing the winner of the linear model (5-parameter) with the winner of the machine learning model (lightGBM), the advanced machine learning model can improve the model accuracy with a margin of 1.1% to 1.7%.

Figure 1: Coefficient of Variation of Root Mean Squared Error (CVRMSE) of the seven algorithms for the three metropolitan areas.



A comparison was made between the model prediction and the ground truth on the test dataset, as depicted in Figure 2. The ground truth was represented by a solid line, while the two linear models and two machine learning models were represented by dashed and dotted lines respectively. Only the two linear and advanced machine learning models were compared to maintain the clarity of the plot.

Figure 2: Model prediction of city-scale daily electricity consumption (GWh) on the test dataset



V. CONCLUSION:

City-level electricity usage is temperature sensitive because heating and cooling are significant energy consumers. As a result of climate change, extreme weather events happen more frequently. A more accurate electricity demand forecast, accounting for extreme weather events and associated load impacts, is needed to enhance the energy security and resilience of the electric grid.

A literature review identified three common approaches to model city-level electricity usage: linear models, machine learning models for time series data (Autoregressive integrated moving average, Recurrent Neural Network/Long Short-Term Memory), and machine learning models for tabular data (neural network-based, decision tree based, and others). In this study, we developed and compared seven data-driven models:

- a five-parameter change-point model
- a Heating Cooling Degree Hour model
- a decomposed time series model implemented by Facebook Prophet
- a Gradient Boosting Trees model implemented by Microsoft lightGBM

The decomposed model has rarely been used in this field; however, lightGBM has been proven as a top performer in city-scale energy demand prediction.

We tested seven models with the city-level (including the city and surrounding rural area) electricity usage data from three metropolitan areas in India: Hyderabad, Mumbai, and Delhi. All the models can predict the city-level electricity demand well, with a CVRMSE of less than 10%. The five-parameter model outperforms the HCDH model. Gradient Boosting Machine is the most accurate model among the seven. The CVRMSE of lightGBM on the test dataset was 6.5% for Mumbai, 4.6% for Hyderabad, and as low as 4.1% for Delhi metropolitan area. Though less accurate than lightGBM, the decomposed time series model gives us a unique chance to decouple and compare the effects of different driving factors (weather-related, yearly and weekly cycle, general trend) of energy demand [56].

VI. REFERENCES:

[1] C. E. Kontokosta and C. Tull, "A data-driven predictive model of city-scale energy use in buildings," Appl. Energy, vol. 197, pp. 303–317, Jul. 2017, doi: 10.1016/j.apenergy.2017.04.005.

[2] O. Deschênes and M. Greenstone, "Climate Change, Mortality, and Adaptation: Evidence from Annual Fluctuations in Weather in the US," Am. Econ. J. Appl. Econ., vol. 3, no. 4, pp. 152–185, Oct. 2011, doi: 10.1257/app.3.4.152.

[3] J. R. Vázquez-Canteli, S. Ulyanin, J. Kämpf, and Z. Nagy, "Fusing TensorFlow with building energy simulation for intelligent energy management in smart cities," Sustain. Cities Soc., vol. 45, pp. 243–257, Feb. 2019, doi: 10.1016/j.scs.2018.11.021.

[4] S. Davoudi, E. Brooks, and A. Mehmood, "Evolutionary Resilience and Strategies for Climate Adaptation," Plan. Pract. Res., vol. 28, no. 3, pp. 307–322, Jun. 2013, doi: 10.1080/02697459.2013.787695.

[5] H. Yi-Ling, M. Hai-Zhen, D. Guang-Tao, and S. Jun, "Influences of Urban Temperature on the Electricity Consumption of Shanghai," Adv. Clim. Change Res., vol. 5, no. 2, pp. 74–80, Jan. 2014, doi: 10.3724/SP.J.1248.2014.074.

[6] J. A. Sathaye, L. L. Dale, P. H. Larsen, and G. A. Fitts, "Estimating impacts of warming temperatures on California's electricity system," Glob. Environ. Change, vol. 23, no. 2, pp. 499–511, Apr. 2013, doi: https://doi.org/10.1016/j.gloenvcha.2012.12.005.

[7] "Joint Response to Governor Newsom Letter August192020.pdf." Accessed: Oct. 01, 2020. [Online]. Available: https://www.cpuc.ca.gov/uploadedFiles/CPUCWebsite/Content/News_Room/NewsUpdates/2020/Joint%20Response%20to%20Governor%20Newsom%20Letter%20August192020.pdf.

[8] M. Carvalho, D. B. de M. Delgado, K. M. de Lima, M. de C. Cancela, C. A. dos Siqueira, and D. L. B. de Souza, "Effects of the COVID-19 pandemic on the Brazilian electricity consumption patterns," Int. J. Energy Res., vol. 45, no. 2, pp. 3358–3364, 2021, doi: https://doi.org/10.1002/er.5877.

[9] A. Bahmanyar, A. Estebsari, and D. Ernst, "The impact of different COVID-19 containment measures on electricity consumption in Europe," Energy Res. Soc. Sci., vol. 68, p. 101683, Oct. 2020, doi: 10.1016/j.erss.2020.101683.

[10] B. Janzen and D. Radulescu, "Electricity Use as a Real-Time Indicator of the Economic Burden of the COVID-19-Related Lockdown: Evidence from Switzerland," CESifo Econ. Stud., vol. 66, no. 4, pp. 303–321, Dec. 2020, doi: 10.1093/cesifo/ifaa010.

[11] H. Lim and Z. J. Zhai, "Review on stochastic modeling methods for building stock energy prediction," Build. Simul., vol. 10, no. 5, pp. 607–624, Oct. 2017, doi: 10.1007/s12273-017-0383-y.

[12] L. G. Swan and V. I. Ugursal, "Modeling of end-use energy consumption in the residential sector: A review of modeling techniques," Renew. Sustain. Energy Rev., vol. 13, no. 8, pp. 1819–1835, Oct. 2009, doi: 10.1016/j.rser.2008.09.033.

[13] V. Bianco, O. Manca, and S. Nardini, "Electricity consumption forecasting in Italy using linear regression models," Energy, vol. 34, no. 9, pp. 1413–1421, Sep. 2009, doi: 10.1016/j.energy.2009.06.034.

[14] L. Shorrock and J. Dunster, "The physically-based model BREHOMES and its use in deriving scenarios for the energy use and carbon dioxide emissions of the UK housing stock," Energy Policy, vol. 25, no. 12, pp. 1027–1037, Oct. 1997, doi: 10.1016/S0301-4215(97)00130-4.

[15] Z. Wang, Z. Zhao, B. Lin, Y. Zhu, and Q. Ouyang, "Residential heating energy consumption modeling through a bottom-up approach for China's Hot Summer–Cold Winter climatic region," Energy Build., vol. 109, pp. 65–74, Dec. 2015, doi: 10.1016/j.enbuild.2015.09.057.

[16] C. E. Kontokosta and C. Tull, "A data-driven predictive model of city-scale energy use in buildings," Appl. Energy, vol. 197, pp. 303–317, Jul. 2017, doi: 10.1016/j.apenergy.2017.04.005.

[17] Y. Chen, T. Hong, and M. A. Piette, "Automatic generation and simulation of urban building energy models based on city datasets for city-scale building retrofit analysis," Appl. Energy, vol. 205, pp. 323– 335, Nov. 2017, doi: 10.1016/j.apenergy.2017.07.128.

[18] N. Abbasabadi and M. Ashayeri, "Urban energy use modeling methods and tools: A review and an outlook," Build. Environ., vol. 161, p. 106270, Aug. 2019, doi: 10.1016/j.buildenv.2019.106270.

[19] M. Lindsey, J. L. Schofer, P. Durango-Cohen, and K. A. Gray, "The effect of residential location on vehicle miles of travel, energy consumption and greenhouse gas emissions: Chicago case study," Transp. Res. Part Transp. Environ., vol. 16, no. 1, pp. 1–9, Jan. 2011, doi: 10.1016/j.trd.2010.08.004.

[20] P. Kuusela, I. Norros, R. Weiss, and T. Sorasalmi, "Practical lognormal framework for household energy consumption modeling," Energy Build., vol. 108, pp. 223–235, Dec. 2015, doi: 10.1016/j.enbuild.2015.09.008.

[21] S. K. Park, H. J. Moon, K. C. Min, C. Hwang, and S. Kim, "Application of a multiple linear regression and an artificial neural network model for the heating performance analysis and hourly prediction of a large-scale ground source heat pump system," Energy Build., vol. 165, pp. 206–215, Apr. 2018, doi: 10.1016/j.enbuild.2018.01.029.

[22] J. D. Olden, M. K. Joy, and R. G. Death, "An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data," Ecol. Model., vol. 178, no. 3, pp. 389– 397, Nov. 2004, doi: 10.1016/j.ecolmodel.2004.03.013.

[23] J. K. Kissock, J. S. Haberl, and D. E. Claridge, "Development of a Toolkit for Calculating Linear, Change-Point Linear and Multiple-Linear Inverse Building Energy Analysis Models, ASHRAE Research Project 1050-RP, Final Report," Energy Systems Laboratory, Texas A&M University, Technical Report, Nov. 2002. Accessed: Jan. 19, 2021. [Online]. Available: https://oaktrust.library.tamu.edu/handle/1969.1/2847.

[24] Y. Zhang, Z. O'Neill, B. Dong, and G. Augenbroe, "Comparisons of inverse modeling approaches for predicting building energy performance," Build. Environ., vol. 86, pp. 177–190, Apr. 2015, doi: 10.1016/j.buildenv.2014.12.023.

[25] Y. Geng, W. Ji, B. Lin, J. Hong, and Y. Zhu, "Building energy performance diagnosis using energy bills and weather data," Energy Build., vol. 172, pp. 181–191, Aug. 2018, doi: 10.1016/j.enbuild.2018.04.047.

[26] H. Li, A. Bekhit, C. Szum, C. Nesler, S. Lisauskas, and S. C. Snyder, "Targeting Building Energy Efficiency Opportunities: An Open-source Analytical & Benchmarking Tool," ASHRAE Trans., vol. 125, pp. 470–478, 2019.

[27] O. Büyükalaca, H. Bulut, and T. Yılmaz, "Analysis of variable-base heating and cooling degree-days for Turkey," Appl. Energy, vol. 69, no. 4, pp. 269–283, Aug. 2001, doi: 10.1016/S0306-2619(01)00017- 4.

[28] S. N. Chandramowli and F. A. Felder, "Impact of climate change on electricity systems and markets – A review of models and forecasts," Sustain. Energy Technol. Assess., vol. 5, pp. 62–74, Mar. 2014, doi: 10.1016/j.seta.2013.11.003.

[29] A. D'Amico, G. Ciulla, D. Panno, and S. Ferrari, "Building energy demand assessment through heating degree days: The importance of a climatic dataset," Appl. Energy, vol. 242, pp. 1285–1306, May 2019, doi: 10.1016/j.apenergy.2019.03.167.

[30] A. Bolattürk, "Optimum insulation thicknesses for building walls with respect to cooling and heating degree-hours in the warmest zone of Turkey," Build. Environ., vol. 43, no. 6, pp. 1055–1064, Jun. 2008, doi: 10.1016/j.buildenv.2007.02.014.

[31] Gh. R. Roshan, A. A. Ghanghermeh, and S. Attia, "Determining new threshold temperatures for cooling and heating degree day index of different climatic zones of Iran," Renew. Energy, vol. 101, pp. 156– 167, Feb. 2017, doi: 10.1016/j.renene.2016.08.053.

[32] "scipy.optimize.curve_fit — SciPy v1.5.2 ReferenceGuide." https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.curve_fit.html (accessed Oct. 01, 2020).

[33] J. G. De Gooijer and R. J. Hyndman, "25 years of time series forecasting," Int. J. Forecast., vol. 22, no. 3, pp. 443–473, Jan. 2006, doi: 10.1016/j.ijforecast.2006.01.001.

[34] E. Erdogdu, "Natural gas demand in Turkey," Appl. Energy, vol. 87, no. 1, pp. 211–219, Jan. 2010, doi: 10.1016/j.apenergy.2009.07.006.

[35] S. Saab, E. Badr, and G. Nasr, "Univariate modeling and forecasting of energy consumption: the case of electricity in Lebanon," Energy, vol. 26, no. 1, pp. 1–14, Jan. 2001, doi: 10.1016/S0360- 5442(00)00049-9.

[36] S. Sp. Pappas, L. Ekonomou, D. Ch. Karamousantas, G. E. Chatzarakis, S. K. Katsikas, and P. Liatsis, "Electricity demand loads modeling using AutoRegressive Moving Average (ARMA) models," Energy, vol. 33, no. 9, pp. 1353–1360, Sep. 2008, doi: 10.1016/j.energy.2008.05.008.

[37] A. J. Conejo, M. A. Plazas, R. Espinola, and A. B. Molina, "Day-ahead electricity price forecasting using the wavelet transform and ARIMA models," IEEE Trans. Power Syst., vol. 20, no. 2, pp. 1035– 1042, May 2005, doi: 10.1109/TPWRS.2005.846054.

[38] L. Suganthi and A. A. Samuel, "Energy models for demand forecasting—A review," Renew. Sustain. Energy Rev., vol. 16, no. 2, pp. 1223–1240, Feb. 2012, doi: 10.1016/j.rser.2011.08.014.

[39] A. Rahman, V. Srikumar, and A. D. Smith, "Predicting electricity consumption for commercial and residential buildings using deep recurrent neural networks," Appl. Energy, vol. 212, pp. 372–385, Feb. 2018, doi: 10.1016/j.apenergy.2017.12.051.

[40] W. Wang, T. Hong, X. Xu, J. Chen, Z. Liu, and N. Xu, "Forecasting district-scale energy dynamics through integrating building network and long short-term memory learning algorithm," Appl. Energy, vol. 248, pp. 217–230, Aug.2019, doi: 10.1016/j.apenergy.2019.04.085.

[41] Z. Wang, T. Hong, and M. A. Piette, "Predicting plug loads with occupant count data through a deep learning approach," Energy, vol. 181, pp. 29–42, Aug. 2019, doi: 10.1016/j.energy.2019.05.138.

[42] "Prophet," Prophet. http://facebook.github.io/prophet/ (accessed Oct. 02, 2020).

[43] S. J. Taylor and B. Letham, "Forecasting at Scale," Am. Stat., vol. 72, no. 1, pp. 37–45, Jan. 2018, doi: 10.1080/00031305.2017.1380080.

[44] J. Asha, S. Rishidas, S. SanthoshKumar, and P. Reena, "Analysis of Temperature Prediction Using Random Forest and Facebook Prophet Algorithms," in Innovative Data Communication Technologies and Application, Cham, 2020, pp. 432–439, doi: 10.1007/978-3-030-38040-3_49.

[45] W.-X. Fang, P.-C. Lan, W.-R. Lin, H.-C. Chang, H.-Y. Chang, and Y.-H. Wang, "Combine Facebook Prophet and LSTM with BPNN Forecasting financial markets: the Morgan Taiwan Index," in 2019 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), Dec. 2019, pp. 1–2, doi: 10.1109/ISPACS48206.2019.8986377.

[46] S. Mahmud, "Bangladesh COVID-19 Daily Cases Time Series Analysis using Facebook Prophet Model," Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 3660368, Jul. 2020. doi: 10.2139/ssrn.3660368.

[47] I. Fernández, C. E. Borges, and Y. K. Penya, "Efficient building load forecasting," in ETFA2011, Sep. 2011, pp. 1–8, doi: 10.1109/ETFA.2011.6059103.

[48] G. K. F. Tso and K. K. W. Yau, "Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks," Energy, vol. 32, no. 9, pp. 1761–1768, Sep. 2007, doi: 10.1016/j.energy.2006.11.010.

[49] J. Roth, A. Bailey, S. Choudhary, and R. K. Jain, "Spatial and Temporal Modeling of Urban Building Energy Consumption Using Machine Learning and Open Data," pp. 459–467, Jun. 2019, doi: 10.1061/9780784482445.059.

[50] Y. Li, Z. Wen, Y. Cao, Y. Tan, D. Sidorov, and D. Panasetsky, "A combined forecasting approach with model self-adjustment for renewable generations and energy loads in smart community," Energy, vol. 129, pp. 216–227, Jun. 2017, doi: 10.1016/j.energy.2017.04.032.

[51] F. H. Al-Qahtani and S. F. Crone, "Multivariate k-nearest neighbour regression for time series data — A novel algorithm for forecasting UK electricity demand," in the 2013 International Joint Conference on Neural Networks (IJCNN), Aug. 2013, pp. 1–8, doi: 10.1109/IJCNN.2013.6706742.

[52] J. A. Fonseca and A. Schlueter, "Integrated model for characterization of spatiotemporal building energy consumption patterns in neighborhoods and city districts," Appl. Energy, vol. 142, pp. 247–265, Mar. 2015, doi: 10.1016/j.apenergy.2014.12.068.

[56] Wang, Z., Hong, T., Li, H. and Piette, M.A., 2021. Predicting City-Scale Daily Electricity Consumption Using Data-Driven Models. Advances in Applied Energy, p.100025.
@article {wang2021predicting,
title= {Predicting City-Scale Daily Electricity Consumption Using Data-Driven Models},
author= {Wang, Zhe and Hong, Tianzhen and Li, Han and Piette, Mary Ann},
journal= {Advances in Applied Energy},
pages= {100025},
year= {2021},
publisher={Elsevier}
}