Predicting Energy Consumption in London Using Smart Meter Data and Machine Learning Models

Hemang Mehta
Departement of CSE
PDEU

Deven Bariya
Department of CSE
PDEU

Dhruvil Patel
Department of CSE
PDEU

Sohma Vyas
Department of CSE
PDEU

***Abstract*—**

This research paper investigates the use of machine learning models to predict energy consumption in London leveraging data from smart meters distributed across the city. The dataset comprises various weather characteristics alongside hourly energy consumption records. Multiple machine learning algorithms were implemented to determine the most accurate model for predicting energy usage. The study involved preprocessing the data, handling missing values, and normalizing features to optimize model performance. Evaluation metrics such as Accuracy, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared were utilized to compare the predictive capabilities of each model. The findings suggest that machine learning models can effectively predict energy consumption based on weather conditions, offering valuable insights for energy management and resource allocation in urban environments like London. Future research could explore ensemble techniques and deep learning models for further improving predictive accuracy.

Keywords*— *smart meters, energy prediction, machine learning

I. Introduction

The rising demand for sustainable energy solutions necessitates a deeper understanding of residential electricity consumption patterns. In this context, smart meters have emerged as a game-changer, providing high-resolution data on household power usage. This research focuses on leveraging this data to predict hourly electricity consumption in London households. Accurately predicting hourly power usage offers significant advantages for both consumers and utility providers. For consumers, it allows for informed energy management practices, leading to cost optimization through targeted reductions in peak demand periods. Utility providers benefit from improved demand forecasting capabilities, enabling them to optimize generation and distribution strategies. This translates to increased grid stability and potentially lower electricity prices for consumers. Traditional methods for predicting electricity consumption often rely on aggregated data and may not capture the intricacies of individual household behavior. This study proposes a novel machine learning model that incorporates smart meter readings alongside key external factors known to influence energy usage. These factors include weather variables like temperature and humidity, designated holidays and weekends, and the inherent hourly fluctuations in power demand. By integrating these diverse elements, the model aims to achieve superior prediction accuracy compared to existing methods. The research further explores the specific influence of individual weather elements on household energy consumption patterns. We aim to elucidate how variations in temperature, humidity, and precipitation levels impact decisions related to energy use, such as the operation of air conditioning and heating systems. This study utilizes a comprehensive dataset provided in the UCI Repository [1]. By evaluating the effectiveness of the proposed model on this data, we aim to shed light on the feasibility of achieving highly accurate hourly power usage predictions. The anticipated results will pave the way for the development of informed energy management strategies that benefit both consumers and utility providers in the London area.

II. Literature Review

Fayaz et al. [2] delve into Deep Extreme Learning Machines (DELM) for residential building energy consumption prediction. By focusing on DELM and comparing it to established techniques like ANNs and ANFIS, they provide specific insights into the potential of this emerging approach for residential energy prediction. This targeted investigation strengthens the understanding of advanced machine learning techniques in this application domain. Dara et al. [3] present a comparative study on various models. Their research delves into the practical application of different techniques, analyzing their performance in predicting electricity consumption. This comparison offers valuable insights into the strengths and limitations of each model, aiding researchers in selecting the most effective approach for their specific forecasting requirements. Srinivasan et al. [4] propose an Extreme Learning Machine (ELM) approach for energy disaggregation. This approach tackles the challenge of identifying individual appliance power consumption from a single overall household meter reading. Their work explores an alternative to established techniques and offers insights into the effectiveness of ELMs for this specific application. This investigation broadens the landscape of potential solutions for accurate energy disaggregation. Luo et al. [5] present a comparative study on machine learning frameworks for multi-objective prediction of various building energy loads. Their work focuses on predicting heating, cooling, lighting demands, and photovoltaic (BIPV) power generation simultaneously. The study investigates three machine learning techniques: Artificial Neural Networks (ANNs), Support Vector Regression (SVR), and Long-Short Term Memory (LSTM) networks. Their findings suggest that a multi-objective approach can efficiently predict multiple building energy aspects while saving computational time. Divina et al. [6] evaluate various time series forecasting techniques for predicting short-term electricity use in smart buildings. They compare traditional statistical methods like ARIMA with machine learning approaches such as Support Vector Regression (SVR) and K-Nearest Neighbors (KNN). Their findings suggest that machine learning models

outperform statistical methods in predicting short-term building energy consumption

III. Methodology

The dataset used contains 14 columns from which there is 1 output variable called 'Energy' and 13 input variables that contain weather data. There were three columns that contained string data type. The dataset was processed using the following steps, 'Compiling the data' where the data considered for this work was the half hourly data provided in a block wise format. It was converted to an hourly format to integrate along with the weather data provided by the repository. 'Filling the empty values', where the empty values were replaced using the mean values from the rest of the column. 'Converting the columns' where the columns with data-type as 'String' were replaced with integer values to make the data viable for the models to work on. 'Deleting rows' where instances where the 'Energy' column had Nan value were dropped. There were 1543 such instances. 'Model implementation' where the following models were implemented to predict the hourly energy consumption,

1. Linear Regressor –

Linear regression is a statistical method used for modeling the relationship between a dependent variable (target) and one or more independent variables (features). It assumes a linear relationship between the variables, meaning that changes in the independent variables result in proportional changes in the dependent variable.

The formula for a simple linear regression model with one independent variable can be expressed as:

$y= mx + b$

Where:

y is the dependent variable (target),

x is the independent variable (feature),

m is the slope of the line (coefficient),

b is the y-intercept.

2. KNN Regressor –

K-Nearest Neighbors (KNN) regression is a non-parametric machine learning algorithm used for regression tasks, where the goal is to predict continuous values. Unlike traditional regression methods, KNN regression makes predictions based on the similarity between data points. When tasked with predicting the output for a new data point, KNN regression identifies the k nearest neighbors from the training dataset using a distance metric (typically Euclidean distance) and then calculates the average of their target values to make the prediction. This approach makes KNN regression robust to different types of data distributions and is particularly useful when dealing with datasets where the relationship between features and target values is not easily modeled by parametric methods. However, KNN regression can be computationally expensive, especially with large datasets, as it requires storing all training data and computing distances for each prediction.

Fig 1. Equation for SVM Regressor

$$\hat{y}_{\text{new}} = \frac{1}{k} \sum_{x_i \in N_k(x_{\text{new}})} y_i$$

where $\hat{y}_{new}$ is the predicted target value for a new data point x_{new} is the number of nearest neighbors, x_i are the training data points, $N_k(x_{new})$is the set of k nearest neighbors of x_{new}, and y_i are the corresponding target values of the nearest neighbors.

3. Random Forest Regressor –

Random Forest Regressor is an ensemble learning algorithm widely used for regression tasks, known for its robustness and high performance. It operates by constructing multiple decision trees during training and outputs the average prediction of the individual trees for regression tasks. Each tree in the forest is built using a random subset of the training data and a random subset of features, ensuring diversity among the trees. During prediction, the algorithm aggregates the predictions of all trees to produce the final output. Random Forest Regressor is effective in handling high-dimensional data, capturing nonlinear relationships, and mitigating overfitting. Its flexibility, scalability, and ability to handle missing values make it a popular choice for various regression applications in practice.

The formula is as shown in equation 2,

Fig 2. Equation for SVM Regressor

$$\hat{y}_{\text{new}} = \frac{1}{N} \sum_{i=1}^{N} f_i(x_{\text{new}})$$

Here, N is the number of trees in the forest, $f_i(x_{new})$ is the prediction of the i_{th} decision tree for the new data point x_{new}.The final prediction is the average of predictions from all individual trees in the forest. This ensemble approach reduces overfitting and variance, making Random Forest Regressor robust and effective for various regression tasks.

4. Ridge Regressor –

Ridge Regressor, a variant of linear regression, is a regularization technique used to mitigate multicollinearity and overfitting in regression models. It introduces an additional term, the L2 penalty term, to the ordinary least squares (OLS) objective function, which penalizes large coefficients, thus shrinking them towards zero. This regularization helps stabilize the

model by reducing variance, making it less sensitive to changes in the training data and improving generalization performance. Ridge Regressor is particularly effective when dealing with datasets with high dimensionality or multicollinearity among the features. It strikes a balance between bias and variance, offering a robust solution for regression tasks in various domains.

Fig 3. Equation for SVM Regressor

$$RSS_{L2} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \lambda \sum_{j=1}^{P} B_j^2$$

Here, Y is the predicted value (dependent variable), X is any predictor (independent variable), B is the regression coefficient attached to that independent variable, and X0 is the value of the dependent variable when the independent variable equals zero (also called the y-intercept).

5. SVM Regressor –

Support Vector Machine (SVM) Regressor is a powerful supervised learning algorithm used for regression tasks. Unlike traditional regression techniques, SVM Regressor aims to find the optimal hyperplane that best fits the data while maximizing the margin between the hyperplane and the data points. This hyperplane is determined by support vectors, which are the data points closest to the hyperplane. The algorithm seeks to minimize the prediction error while penalizing points that fall outside a specified margin. SVM Regressor can effectively handle high-dimensional data and is particularly useful when dealing with nonlinear relationships between features and target values, thanks to its ability to use different kernel functions. Additionally, SVM Regressor is robust to overfitting and performs well even with small to medium-sized datasets, making it a popular choice for regression tasks in various domains.

Fig 4. Equation for SVM Regressor

$$\hat{y}_{\text{new}} = \sum_{i=1}^{n_{\text{sv}}} \alpha_i y_i K(x_i, x_{\text{new}}) + b$$

where y^new is the predicted output for a new data point xnew, nsv is the number of support vectors, alpha_i are the Lagrange multipliers obtained during training, yi are the corresponding target values of the support vectors, K(xi, xnew) is the kernel function which measures the similarity between xi and xnew and b is the bias term.

6. XGBoost Regressor –

XGBoost (eXtreme Gradient Boosting) is a popular machine learning algorithm for regression and classification tasks. It belongs to the ensemble learning family, specifically boosting algorithms, which build predictive models by combining the outputs of several weaker models (often decision trees) to improve overall accuracy and generalization.

Fig 5. Equation for SVM Regressor

$$\text{Obj} = \sum_{i=1}^{n} \mathcal{L}(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Here,

L = Loss function

yi = actual target value

y^i = predicted target value

Omega = regularization term

K = Total number of trees in ensemble

IV. Result

Parameters such as Mean Square Error and Accuracy were determined for each model to predict their efficiency as shown in table (1). Mean squared error (MSE) is a measure of the average squared difference between the actual and predicted values in a regression model. It quantifies the overall accuracy of the model's predictions by penalizing larger errors more than smaller ones. The coefficient of determination, often referred to as (R-squared) score, is a statistical measure that represents the proportion of variance in the dependent variable that is predictable from the independent variables in a regression model. It ranges from 0 to 1, with higher values indicating a better fit of the model to the data. In simpler terms, score quantifies the goodness of fit of the regression model, showing how well the model explains the variability in the target variable based on the predictors used.

Table 1. MSE and R2 score for each model

	LR	RR	RF	KNN	SVM	XGB
MSE	946076	946076	1.07	653636	1.0190	568193
R2	0.085	0.085	0.473	0.368	0.077	0.4506

To maintain visibility and readability, the last 100 points of the testing dataset and the predicted dataset were plotted on the graph. The graphs for their corresponding models are shown in figures 6 to 11.

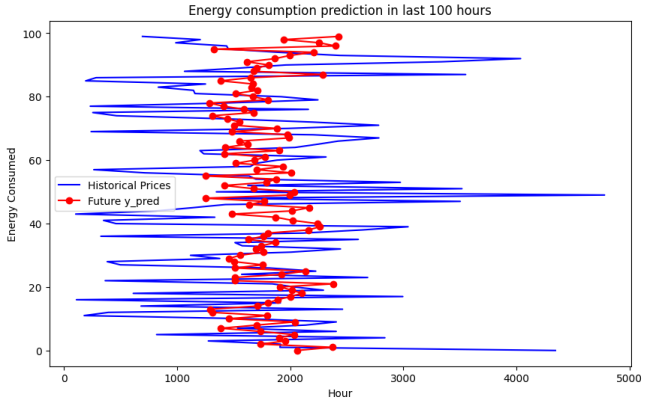

Fig 6. Graph for testing and predicted dataset through Linear Regression

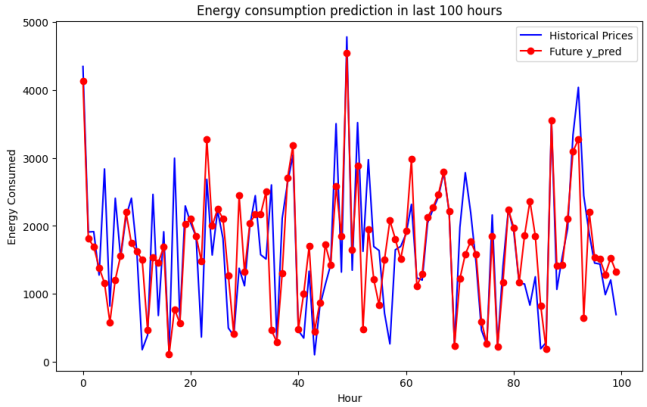

Fig 7. Graph for testing and predicted dataset through KNN

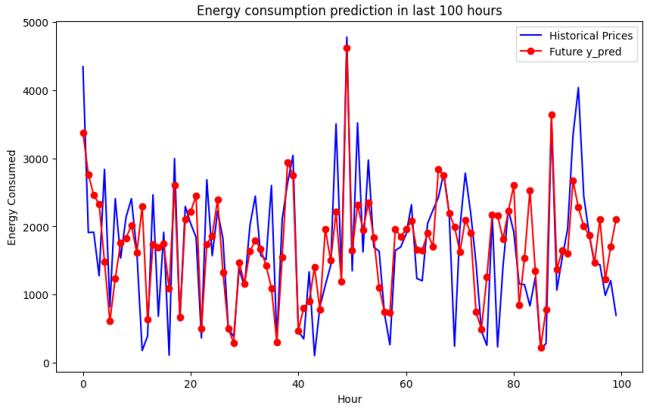

Fig 8. Graph for testing and predicted dataset through Random Forest Classifier

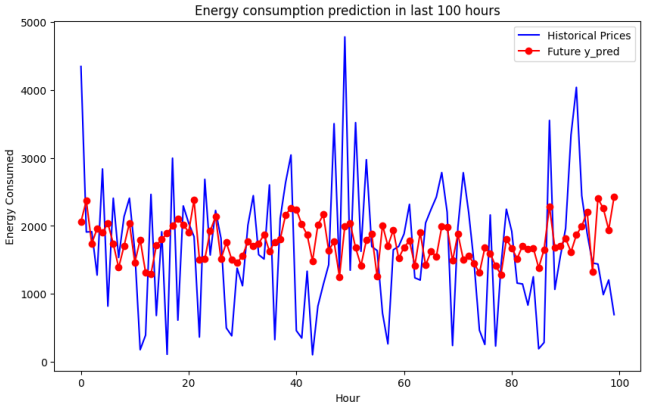

Fig 9. Graph for testing and predicted dataset through Ridge Regressor

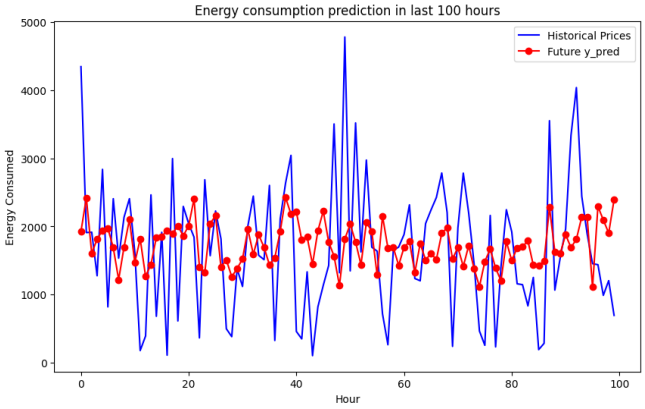

Fig 10. Graph for testing and predicted dataset through SVM Regressor

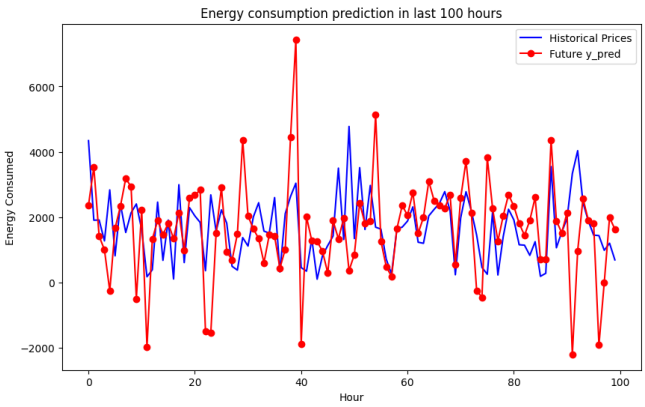

Fig 11. Graph for testing and predicted dataset through XGBoost Regressor

V. Conclusion

This research investigated the feasibility of employing machine learning models to predict energy consumption in London using real-time weather data collected via smart meters. The exploration implemented various machine learning algorithms, with the Random Forest Regressor achieving the most

promising results, reflected by an R-squared value of 0.473. This achievement indicates a moderate correlation between the predicted and actual energy consumption values. This work presents a significant advancement towards establishing real-time energy consumption prediction models for urban areas. By leveraging weather data and smart meter technology, such models can empower stakeholders across the energy sector. Utility companies can gain valuable insights to optimize energy production and distribution based on predicted demand fluctuations. This can lead to improved grid stability and potentially lower energy costs for consumers. Furthermore, real-time prediction empowers individual consumers to make informed decisions about their energy usage, potentially promoting energy conservation efforts within the city.

VI. References

1. https://www.kaggle.com/datasets/jeanmidev/smart-meters-in-london
2. Fayaz, M.; Kim, D. A Prediction Methodology of Energy Consumption Based on Deep Extreme Learning Machine and Comparative Analysis in Residential Buildings. *Electronics* **2018**, *7*, 222. https://doi.org/10.3390/electronics7100222
3. Lee, M.H.L.; Ser, Y.C.; Selvachandran, G.; Thong, P.H.; Cuong, L.; Son, L.H.; Tuan, N.T.; Gerogiannis, V.C. A Comparative Study of Forecasting Electricity Consumption Using Machine Learning Models. *Mathematics* **2022**, *10*, 1329. https://doi.org/10.3390/math10081329
4. Salerno, V.M.; Rabbeni, G. An Extreme Learning Machine Approach to Effective Energy Disaggregation. *Electronics2018,7,*235.https://doi.org/10.3390/electronics7100235
5. X.J. Luo, Lukumon O. Oyedele, Anuoluwapo O. Ajayi, Olugbenga O. Akinade,'Comparative study of machine learning-based multi-objective prediction framework for multiple building energy loads',Sustainable Cities and Society, Volume 61, 2020, 102283, ISSN 2210-6707, https://doi.org/10.1016/j.scs.2020.102283.
6. Divina, F.; García Torres, M.; Goméz Vela, F.A.; Vázquez Noguera, J.L. A Comparative Study of Time Series Forecasting Methods for Short Term Electric Energy Consumption Prediction in Smart Buildings. *Energies* **2019**, *12*, 1934. https://doi.org/10.3390/en12101934