Predicting Household electricity Consumption with Machine Learning and Smart Meter Data

Houda KHELIFI[1] and Sofiane KHALFALLAH[2] and Hela LTIFI[3]

***Abstract*— Predicting energy consumption plays a crucial role in promoting energy conservation and reducing power generation costs. Recent research indicates a growing interest in utilizing machine learning algorithms for predicting energy consumption in households. This research utilizes a massive dataset of several millions, containing electricity consumption records of residential households in Uruguay (mostly in Montevideo). In this study we examine the utilization of multiple machine learning techniques such as linear regression, random forest, extra-trees regressor, lasso regression, xg boost, and ridge regression to predict household electricity demand. These models were trained and evaluated using historical electricity usage data. In order to evaluate these models, the coefficient of determination (R squared) metric is employed. Tree-based algorithms, including random forest, extra trees regressor, and xgboost, achieve the best results. Among them, random forest demonstrates the highest performance.**

Index Terms—Massive dataset, Linear Regression, Random Forest, Extra-Trees Regressor, Lasso Regression, XG Boost, Ridge Regression.

I. INTRODUCTION

The academic study of energy consumption has gained significant attention in recent decades. Energy-related concerns are crucial for ensuring societal security and prosperity, as they directly impact both economic stability and environmental sustainability [1].

Energy consumption continues to rise daily, with significant waste often resulting from residents negligence and lack of awareness regarding peak hours. The problem gets worse when energy production fails to anticipate future demand, leading to inefficiencies [2]. Forecasting energy consumption is crucial for ensuring a reliable power supply. To maintain consistent and dependable electricity availability, it is essential to anticipate energy usage in advance. This is especially crucial due to the concurrent energy demand that arises during power plant operations [3].

The classification of load forecasting techniques depends on both the forecasting horizon and the forecasting region [4]. The forecasting horizon is the expected load demand over a specific time period. It is categorised into three main types. Short-term load forecasting covers a few hours or days. Medium-term load forecasting spans weeks or months. Long-term load forecasting predicts demand over months or years [4] [18]. The focus of the paper is on short-term forecasting, with data recorded at fifteen-minute intervals. Regression models, machine learning techniques and time-series forecasting are among the most commonly utilized methods in this field. The study specifically examines the performance of various regression models, including random forest, xgboost, extra trees regressor, lasso regression, ridge regression and linear regression.

The dataset used in this work contains electricity consumption records from residential households in Uruguay, mostly in Montevideo. It was designed to analyze customer behavior and identify energy consumption patterns [5]. The dataset includes approximately 59 millions records.

The remainder of this research work is organized as follows: section 2 reviews related literature on energy consumption. Section 3 outlines the methodologies employed in the study. Section 4 presents the experimental results. Finally, section 5 provides the conclusions of the paper and some perspectives.

II. LITERATURE REVIEW

A review of relevant publications has been made to predict energy consumption and improve efficiency. The literature published on energy consumption prediction has used machine learning methods. Machine learning can enhance energy efficiency by analyzing large amounts of data from energy systems. It plays a crucial role in identifying patterns and trends, while making accurate predictions about energy consumption. These predictions can help in optimizing energy usage, identifying areas of wastage, and suggesting energy-saving measures.

In [6] the authors used smart meter data from the Republic of Ireland to enhance the understanding of household electricity consumption. They applied generalized additive models to assess the impact of factors such as temperature, day of the week, and month of year on energy usage. The researchers utilized K-means clustering to identify intra-day load profiles. Additionally, in this study, the authors used k-medoids clustering to generate annual load profiles. Various machine learning models were employed to predict yearly electricity consumption based on an analysis of smart meter electricity data and household level characteristics. These models include deep neural networks, random forest, gradient boosting machine, and elastic net. The dataset used in this study consists of smart meter electricity consumption data recorded at 30-minute intervals.

The main drawback of this research work is that it does not employ data with finer granularity. In fact the use of 15-minute intervals, could provide more detailed insights and

[1]Houda KHELIFI is with National School of Engineers of Gabes, University of Gabes, Tunisia houda.khlifi@gmail.com

[2]Sofiane KHALFALLAH is with Centre's Computer Science Research Laboratory (PRINCE LAB), University of Sousse, 4002, Sousse, Tunisia sofiane.khalfallah@gmail.com

[3]Hela LTIFI is with Research Groups in Intelligent Machines, Sfax, Tunisia hela.ltifi@gmail.com

potentially improve prediction accuracy.

The authors in reference [7] proposed an ensemble of prediction model to predict the building hourly electricity consumption. Which combines the Gated Recurrent Unit (GRU) model with a logarithmic electricity consumption gravity model (LE_GRA). This research work used a model called the Gravity-Gated Recurrent Unit Electricity Consumption model (GRA_GRU). In this study authors used two years (17,520 hours) of electricity consumption data with three factors such as: temperature, humidity, and wind speed. The authors demonstrate that the GRA_GRU model possesses impressive prediction and generalization capabilities.

This research work has certain limitations, primarily arising from the dataset's small size. Furthermore, analyzing data at a higher temporal resolution (e.g, 15-minute intervals of electricity usage), could provide valuable insights and potentially enhance prediction accuracy.

In [8], the authors evaluated and compared several models for short-term load forecasting. These models include auto-regressive integrated moving average, multiple linear regression, recursive partitioning, regression trees, conditional inference trees with bootstrap aggregating, and random forest. The researchers in [8] utilized a one-year sample of electricity load data from a residential area for training and testing. They concluded that the random forest model achieved the lowest error and delivered the most accurate forecasting results.

The previous study has several limitations, including the use of dataset with a limited number of records. Furthermore, the dataset consists of hourly electricity load data. Incorporating finer-grained data, such as 15-minute intervals, could potentially improve accuracy and provide deeper insights into the analysis.

The authors in reference [9] explored four regression models such as: multinomial regression, ridge regression, lasso regression, and polynomial regression. These models were used to predict household energy consumption data. The researchers in this study included four features (global active power, global reactive power, voltage and global intensity) to analyze their influence on prediction results. But only one feature (global active power factor) was tested in the experiment. They concluded that polynomial regression outperformed multinomial regression, ridge regression and lasso regression in terms of prediction accuracy.

The study's limitation lies in the use of a dataset containing only 2 millions records. This limitation may not fully capture the variability or complexity of larger dataset. Consequently, this may restrict the generalization of the results.

Reference [15] explores the implementation of various machine learning algorithms to predict building electricity consumption. These include linear regression, lasso regression, ridge regression, elastic net regression, random forest regression, extra trees, support vector regression and decision tree. Based on their analysis, the authors concluded that lasso regression outperforms support vector regression, making it the most effective model in their study.

The previous study is constrained by the use of a dataset of only one million records, which may reduce the reliability of the results and hinder their applicability to larger-scale analyses.

Our study builds upon existing research in electricity consumption forecasting by applying and evaluating a variety of machine learning models. These models include linear regression, ridge regression, lasso regression, random forest, extra trees regressor and xgboost. Unlike previous works that often rely on smaller dataset. This study leverages a large-scale dataset comprising 59 millions records of smart meter electricity consumption. Additionally, it incorporates fine granularity data with 15-minute intervals, providing a more detailed and precise analysis of consumption patterns. The combination of diverse machine learning models improves the robustness and accuracy of forecasting results. Moreover, the use of massive dataset improves prediction accuracy and generalization by capturing complex patterns and trends. As a result, this leads to more reliable and precise forecasts.

III. MODELLING

A. *Coefficient of Determination*

The coefficient of determination, denoted as R^2 (pronounced R squared), is used to evaluate the performance of the models. It measures the proportion of variance in the target variable that can be explained by the given features. In practical terms, it indicates how well the model fits the data [9]. The coefficient of determination can be expressed numerically using the following equation:

$$R^2 = \frac{\mathrm{SS}_{\mathrm{Regression}}}{\mathrm{SS}_{\mathrm{Total}}} \tag{1}$$

Where, $\mathrm{SS}_{\mathrm{Regression}}$ is the sum of the square of the residual error and $\mathrm{SS}_{\mathrm{Total}}$ is the total sum of the error.

B. *Predicting Models*

Time series analysis offers the opportunity to forecast future values based on the past data. It is widely used to predict trends in various fields such as economics, weather, energy consumption and capacity planning, etc. As the term implies, time series analysis data with time-based data (e.g., years, days, hours and minutes) to uncover hidden patterns and insights. The objective is to understand and measure the frequently unpredictable behavior of systems. As flow, we employ the many several predictive models:

1) Random Forest

A supervised learning algorithm, builds multiple decision trees during training and averages their predictions for regression tasks. Using bagging techniques, it creates an ensemble of trees, offering advantages such as anomaly detection, feature importance identification, trend discovery, and insightful visualizations [11]. Random forest effectively addresses the overfitting problem of decision trees by using the concept of stochastic discrimination in classification. The model comprises multiple decision trees, where input values are passed through to generate predictions. Random

Forest, has ability to handle continuous datasets, delivers high accuracy for non-linear data with fewer features [10].

2) Extra-Trees

A supervised learning algorithm, builds multiple decision trees by randomly selecting splits during training and averages their predictions for regression tasks. Unlike random forest, it increases randomness by choosing split thresholds independently of the data. Improving performance on noisy datasets. The key benefits include robust handling of outliers, feature importance estimation, and efficient processing of large datasets [13] [1].

3) XG Boost

A popular gradient boosting algorithm, is widely used for regression tasks. It builds a sequence of decision trees, combining their predictions to minimize errors between observed and predicted values. Xgboost effectively reduces overfitting and provides feature importance metrics by integrating regularization techniques. The process includes data splitting, model initialization, training, hyperparameter tuning and generating predictions for new data [12].

4) Linear Regression

Is a fundamental and widely used predictive model. Its accuracy in complex scenarios depends on incorporating sufficient descriptive variables. This method represents the relationship between two data elements as a linear equation with coefficients that quantify these relationships. The key benefits include robust handling of outliers, feature importance estimation and efficient processing of large datasets [14]. The linear regression equation is expressed as:

$$y = \beta_0 + \beta_1 x + \epsilon \tag{2}$$

where y is the response variable, x is the predictor variable, β_0 and β_1 are the regression coefficients and ϵ is an error to account for the discrepancy between predicted data and the observed data.

5) Ridge Regression

Ridge regression, also known as Tikhonov regularization, is a technique for building simplified models, particularly when the number of predictor variables exceeds the number of observations. It is especially effective in addressing multicollinearity, where predictor variables are highly correlated. Ridge regression introduces a regularization term to the cost function, helping to prevent overfitting and improve model generalization [15]. The regularization term is expressed as:

$$R(w) = \lambda \sum_{j=1}^{p} w_j^2 \tag{3}$$

Where $R(w)$ is the regularization term, w is the feature weight, λ is the regularization parameter and p is the number of variables.

6) Lasso Regression

Least Absolute Shrinkage and Selection Operator (LASSO) is a regularized linear regression technique that uses a regularization term ($L1$, see equation 4) to perform both variable selection and regularization. It adds a penalty term to the loss function, which encourages the model to shrink the coefficients of less important features to zero, effectively removing them from the model [16] [17].The cost function of Lasso regression is given by:

$$F(\theta) = \text{MSE}(\theta) + \alpha \cdot \sum_{i=1}^{n} \theta_i \tag{4}$$

Where MSE is the Mean Squared Error, which measures the difference between the predicted values and the actual values, α is the regularization parameter and $\sum_{i=1}^{n} \theta_i$ is the L1 regularization term.

C. Methodology

The process of regression models follows a structured workflow to ensure accurate predictions and robust performance. The process starts with the original data, which is then preprocessed to address missing values, outliers, and categorical variables. The dataset is then split into training and testing sets using techniques like train_test_split. To improve model performance, standardization is applied to scale numerical features, ensuring that all variables have a comparable range. The model is then fitted to the training data, learning the underlying patterns. Following this, the model is instantiated with the final configuration to be used for prediction. Finally, the model is evaluated using performance metrics such as R squared, root mean squared error or mean absolute error to assess accuracy and generalization. The following steps are shown in fig. 1.

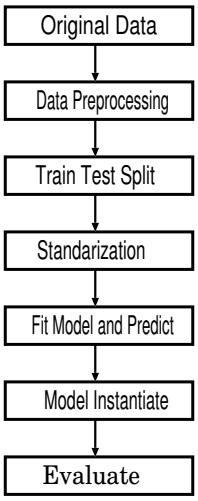

Fig. 1: *Flow chart for modeling of machine learning algorithm*

IV. IMPLEMENTATION AND EXPERIMENT RESULTS

A. *Data*

The dataset used in this study captures the daily energy consumption of households in Uruguay. It comprises total consumption data obtained from smart meters installed across 110,953 households. The data was recorded at fifteen-minute intervals, spanning a monitoring period from $1st$ January 2019 to 3^{rd} November 2020 [5]. This dataset, sourced from the Figshare website[1], includes approximately 59,823,343 records. Each record is structured with three key columns: id (representing the household identifier), datetime (datetime of the record) and value (measured energy consumption). The dataset is organized in a tabular format where each row corresponds to a unique observation, while columns represent distinct variables. This comprehensive dataset provides a rich foundation for analyzing energy consumption patterns and training predictive models. The analysis was conducted using Kaggle Notebooks, a cloud-based platform that provides an integrated environment for data science projects. Python is used as the programming language for the implementation, valued for its versatility and the breadth of its libraries tailored to data analysis and machine learning tasks. The key libraries witch are utilized included Pandas for efficient data manipulation and cleaning. NumPy is used for performing numerical computations. Scikit-learn (sklearn) for implementing machine learning algorithms.

B. *Experiment Results*

To achieve a reliable comparison, the models were assessed based on their R squared values across both the training and test dataset. Furthermore, we analyzed the training time and computational efficiency to assess the practical feasibility of each algorithm for real-world deployment. The evaluation results, presented in Table I, highlight the R squared and training times for each algorithm. Linear regression, lasso, and ridge achieved 0.13% of the variance (R^2-Score = 0.0013) on the test dataset, whereas random forest, xgboost, and extra-trees demonstrated significantly better performance, with 24% of the variance R^2-Score = 0.24). The training time of the random forest model is significantly higher compared to other regression algorithms, reaching 5721.1 seconds. This computational cost is notably greater than that of linear regression, lasso, and ridge regression. These models are inherently more efficient due to their simpler mathematical formulations. They rely on closed-form solutions or iterative optimization techniques. In our experiments on the Kaggle platform, the training process utilized 100% of the available CPU resources and consumed approximately 9 GB of RAM. While exact hardware specifications (CPU, GPU) are not directly accessible. The excessive memory usage and prolonged training time are attributed to the ensemble nature of random forest. It constructs multiple decision trees and performs intensive computations during both training and validation. Despite its high training cost, random forest achieved the best performance in terms of R squared, demonstrating its ability to capture complex patterns in the dataset more effectively than linear models.

[1]Link: https://figshare.com/articles/dataset/THC-preprocessed-files_tar_gzx

TABLE I: *Comparison of Algorithms*

Algorithm	R^2-Score (Train)	R^2-Score (Test)	Training Time (s)
Linear Regression	0.0013607	0.001362	9.7
Ridge Regression	0.0013607	0.0013629	2.1
Lasso Regression	0.0013606	0.0013627	3.4
Random Forest	0.2432762	0.2415318	5721.1
Extra-Trees Regressor	0.2432764	0.2415317	3756.05
XGBoost	0.2432458	0.2415317	185.8

For all the figures used in this section, only the first 1000 examples are selected for visualization due to the large volume of output data. In fig. 2 the dark line shows the predicted values obtained using the linear regression algorithm. While the other represents the actual target values. The graph reveals that the model struggled to deliver accurate predictions.

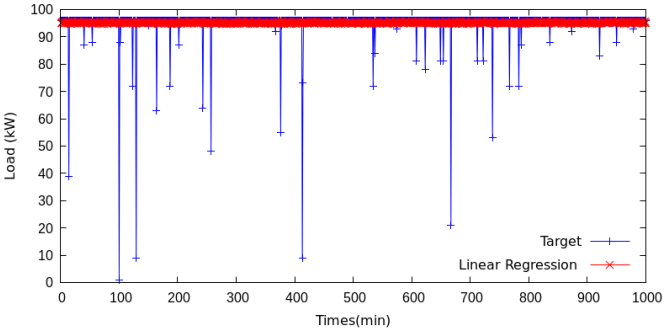

Fig. 2: *Energy Prediction of Linear Regression*

Fig. 3 presents the energy prediction results of the random forest model. One curve represents the actual target values, while the other illustrates the model's predictions. The results indicate that the random forest model performed better compared to the linear model.

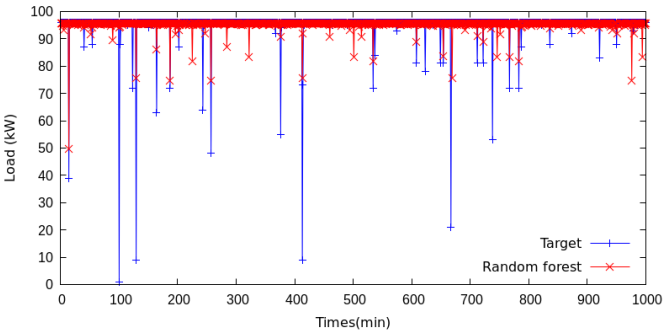

Fig. 3: *Energy Prediction of Random Forest*

The graphs in fig. 3, fig. 4 and fig. 5 show that the tree-based algorithms (random forest, exta trees and xgboost) provide nearly identical predictions. This demonstrates their comparable performance.

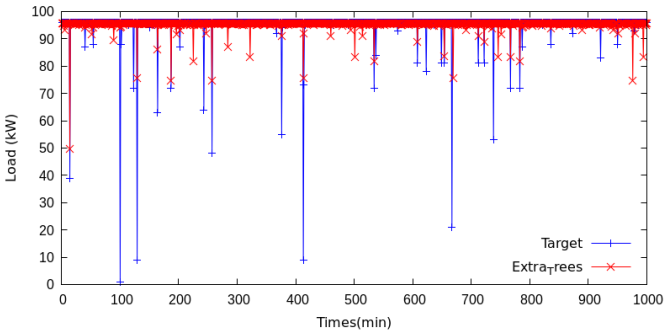

Fig. 4: *Energy Prediction of Extra Trees.*

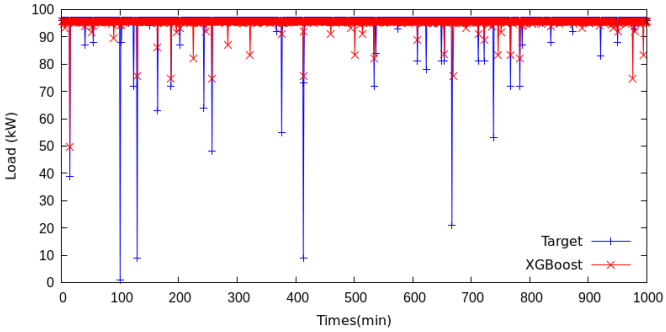

Fig. 5: *Energy Prediction of XGBoost.*

V. CONCLUSION AND FUTURE WORK

In this study, we explore the use of various machine learning algorithms for predicting energy consumption. These include linear regression, lasso regression, ridge regression, random forest, xgboost, and extra trees regressor. The analysis is conducted on a large dataset derived from household energy consumption in Uruguay, collected via smart meters. Our findings highlight that tree-based algorithms, specifically, random forest, xgboost and extra tree regressor, outperformed linear models, achieving 24% of the variance (R^2-Score = 0.24). Among these, random forest demonstrated slightly better performance.

Tree-based models like random forest, extra trees and xgboost are effective for forecasting household energy use due to their ability to model complex, nonlinear patterns with minimal preprocessing.

Future work could improve the model by including external factors and studying the effect of different sampling intervals (15 vs. 30 minutes). Testing existing methods on the same data would help assess generalizability, and applying ensemble models to other energy systems could validate their robustness. Finally, combining R^2 with root mean squared error and mean absolute error ensures a more reliable evaluation.

REFERENCES

[1] S.Amiri, N. Mostafavi, E.Lee, and S. Hoque, Machine learning approaches for predicting household transportation energy use. City and Environment Interactions (CEI), vol. 7, Elsevier 2020, pp. 100044.

[2] A. Nazir, A. Shaikh, A. Shah and A. Khalil, Forecasting energy consumption demand of customers in smart grid using Temporal Fusion Transformer (TFT). Results in Engineering (RE), vol. 1, Elsevier 2023 pp.100888.

[3] L.Saoud, H.Almarzouki and R.hussein, Household Energy Consumption Prediction Using the Stationary Wavelet Transform and Transformers. IEEE Access (2022), vol.10, pp. 5171-5183.

[4] A.Muzumdar, C.Modi and C.Vyjayanthi, An Efficient Regional Short-Term Load Forecasting Model for Smart Grid Energy Management. Annual Conference on Industrial Electronics Society (IECON), IEEE (2021), pp. 2089-2094.

[5] J. Chavat, S.Nesmachnow, J.Graneri AND G.Alvez, ECD-UY, detailed household electricity consumption dataset of Uruguay. Scientific Data (SD), Nature Research (2022).

[6] Z.Guo, Jesse R. O'Hanley and S.Gibson, Predicting residential electricity consumption patterns based on smart meter and household data: A case study from the Republic of Ireland. Utilities Policy (UP), vol.79, Elsevier (2022).

[7] S.Shan, B.Cao and Z.Wu, Forecasting the Short-Term Electricity Consumption of Building Using a Novel Ensemble Model. IEEE Access (2019),vol 7,pp 88093-88106.

[8] A.Kapoor and A.Sharma, A Comparison of Short-Term Load Forecasting Techniques. Innovative Smart Grid Technologies (ISGT), IEEE (2018), pp. 1189-1194.

[9] L.Krishna, A.Kuppusamy and S.Saint Akadiri, Prediction of electrical power consumption in the household: fresh evidence from machine learning approach. Energy Efficiency (EF), Springer (2023), vol. 77,pp. 1-7.

[10] L.Raju, Vishal E, Vishwaraj V and Vimalan K M, Application of Machine Learning Algorithms for Short term Load Prediction of Smart grid. International Conference on Smart Electronics and Communication (ICOSEC), IEEE (2020), pp. 371-376.

[11] H.Wang, Y.Liu, B.Zhou, C.Li, G.Cao, N.Voropai and E.Barakhtenko, Taxonomy research of artificial intelligence for deterministic solar power forecasting. Energy Conversion and Management (ECM), Elsevier (2020), vol. 214, pp. 112909.

[12] I.Lahsen Cherif and A.Kortebi, On using eXtreme Gradient Boosting (XGBoost) Machine Learning algorithm for Home Network Traffic Classification". Wireless Days (WD), IEEE (2019).

[13] V.John, Z.Liu, C.Guo, S.Mita and K.Kidono, Real-Time Lane Estimation Using Deep Features and Extra Trees Regression. International Publishing Switzerland (IPS), Springer (2016), pp. 721–733.

[14] M.Kim, Y.Kim, J.Srebric, Predictions of electricity consumption in a campus building using occupant rates and weather elements with sensitivity analysis: Artificial neural network vs. linear regression. Sustainable Cities and Society (SCS), Elsevier (2020), vol. 64, pp. 102385.

[15] M.Kaur, S.Panwar, A.Joshi and K.Gupta, Residential Electricity Demand Prediction using Machine Learning. International Semantic Intelligence Conference (ISIC), Springer (2021), pp. 331-340.

[16] F.Al-Obeidat, B.Spencer, O.Alfandi, Consistently accurate forecasts of temperature within buildings from sensor data using ridge and lasso regression. Future Generation Computer Systems (FGCS), Elsevier (2018), vol. 110, pp. 382-392.

[17] X.Xiong, Y.Wei, The Analysis and Predication of Energy Use in Smart Homes Based on Machines Learning. International Conference on Computing and Data Scienc (CDS), IEEE (2020).

[18] I. Ali, A. Agga, M. Ouassaid, M. Maaroufi, A. Elrashidi and H. Kotb, Predicting short-term energy usage in a smart home using hybrid deep learning models. Frontiers in Energy Research (FER), Frontiers (2024), vol. 12, id. 1323357.