Proceedings of the 4th International Conference on Computing and Data Science (CONF-CDS 2022)
DOI: 10.54254/2755-2721/2/20220648

Machine Learning for Exploration of the Factors Affecting the Building Energy Usage: A Case Study

Sumaira Asif[1], Tanzila Saba[2,*]

[1] Artificial Intelligence and Data Analytics Lab (AIDA), CCIS
[2] Prince Sultan University, Riyadh, Saudi Arabia

*Corresponding author's e-mail: tsaba@psu.edu.sa

Abstract. This paper presents a thorough exploratory analysis of one of the data research tracks involving the build-ing energy usage data of different cities of France. To gain a deep understanding of the dataset and the various attributes that comprise the data, we performed an exploratory data analysis and used data visual-ization to study the different variables and the correlation that exists between them. This was done to de-termine which factors really contribute to having a huge effect on the energy (gas) consumption leading to climate change. In our study, we also predicted the level of its consumption using a Decision-tree based algorithm and Light GBM, which was found to give satisfactory results.

Keywords: Machine Learning, Energy, Prediction, Technological Development.

1. Introduction

Climate Change means abnormal or unexpected changes in the temperatures and the weather across the world. Climate change is aggravating at an unprecedented scale, it is having huge adverse effects on the living species globally. Energy consumption has enormous effects on Climate change. Energy Consumption means all the energy that is utilized to do a particular action. As the global average temperatures increase due to climate change, people resort to using more energy, particularly electricity. One of the best examples of energy consumption can be a household where energy consumption takes place in the form of electricity, gas etc.[1,2]

The building sector is the most critical factor affecting greenhouse gas emissions, contributing one-third of energy-related EU emissions. The combination of fossil fuel energy and the production of electricity and heat across the buildings are influencing Greenhouse gas emissions [3]. The huge amount of energy utilized in buildings has given rise to many environmental problems. This also had a detrimental impact on the human life. Predicting building energy usage is one of the most effective energy conservation methods, resulting in decisions to decrease energy usage. The energy-efficient buildings can be built, reducing the total energy consumed in new buildings.

Machine learning (ML), a subset of AI uses algorithms that work on historical data and make accurate predictions [4,5]. Machine Learning is effectively being applied to real-life climate change problems such as predicting Carbon-dioxide or greenhouse gas emissions, sea level rise, global ice loss, regions more prone to wildfires risk and the impact of climate change across the globe. At the intersection of climate change and machine learning lies the initiative climate change AI, taken worldwide to study the

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2/20220648

impact of AI on climate change. Therefore, ML can be applied to predict the energy consumption in buildings [6]. This research focused on using machine learning to predict the energy (gas) consumption in the buildings across different cities of France and we explored and analyzed the data set in depth to discover correlations between features that may affect the energy consumption.

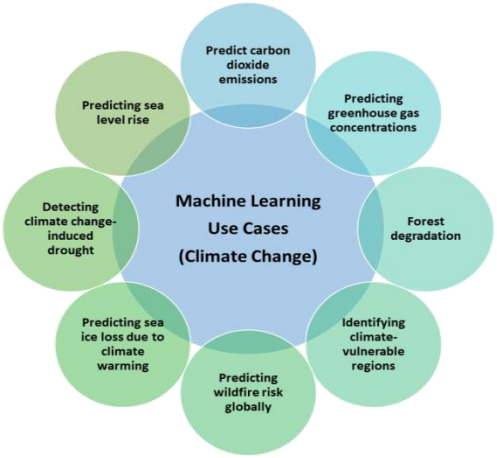

Figure 1. Applications of Machine Learning in Climate Change.

2. Related Work

Pathak et al., [7] proposed two methods to model the gas consumption i.e., Generalized Additive Models (GAM) and Long Short-Term Memory (LSTM). This study utilized building datasets from two different continents. They analyzed significant features present in the dataset and their impact on forecasting gas consumption. GAM and LSTM techniques were implemented and compared with other techniques and it was shown that LSTM gave the best performance compared to GAM and other methods. But GAM gave better interpretable results for building management systems.

De Keijzer et al., [8], employed different machine learning algorithms which were compared to predict the gas consumption forecasting. The dataset used for this study was the gas consumption of 52 housings taken over a period of nine months. Deep Neural Networks (DNN) gave high performance with a 50.1 percent Mean Absolute Percentage Error (MAPE) on a one-hour resolution. Multivariate Linear Regression (MVLR) performed better than other techniques, 20.1 percent MAPE for daily resolutions and 17.0 percent MAPE for weekly resolutions.

In another study reported in [9], machine learning techniques such as Artificial Neural Network (ANN), Gradient Boosting (GB), Deep Neural Network (DNN), Random Forest (RF), Stacking, K Nearest Neighbor (KNN), Support Vector Machine (SVM), Decision tree (DT) and Linear Regression (LR) were used to predict the building energy consumption yearly using a dataset of residential buildings. DNN was found to give the best predictive performance for energy use at the early design phase. Based on this, building designers can make better decisions before constructing them.

In [10], researchers proposed machine learning techniques to improve the accuracy of predicting energy consumption. This work analyses the overall prediction using error curve learning and a hybrid model using consumption data of South Korea. A hybrid model for predicting was built using Catboost, Xgboost, and Multi-layer perceptron algorithms. Mean error was found to be 2.78% on weekdays, 2.79% in case of weekends & 4.28% on special days.

In [11], authors used machine learning techniques like Neural Networks (NN) and regression algorithms such as MLR and Random Forest to build a model for future natural gas consumption using dataset comprising of the data of various small cities in Poland. The results of this research work showed that compared to all the algorithms, the algorithm that performed the best was found to be Random Forest (RF).

DOI: 10.54254/2755-2721/2/20220648

In current research [12], five regression algorithms were used to build models to forecast the fuel consumption rate of gasoline vehicles in China. The various factors that were considered were vehicle, environment & driving behavior factors. RF algorithm performed the best comparatively with a mean absolute error of 0.630 L/100 km. They also found out that the most significant features affecting fuel consumption among the 25 factors present in the study. As per the results, the most significant factors are brake and accelerator habits, engine power, and fuel economy consciousness of vehicle owners in sequence.

3. Exploratory Data Analysis

3.1. Dataset Description

The dataset contains more than 50k observations, including individual building level gas use in 2019 in different cities across France. It also includes individual buildings footprints and other variables indicating some information about the local weather data. We have attributes such as the city name, geometry of the building, no. of flats connected to the gas network, details of the buildings like height, age, floors etc. This dataset was built by combining relevant datasets and preprocessing techniques were applied on these datasets.

3.2. Descriptive Statistics

Table 1 provides a summary of the descriptive statistics of the independent variables of our dataset that are used for the prediction of the consumption level. The average count of records is 55K, the maximum gas consumption is 200 and the minimum is 0.3. The average for building height, floors, wall, and roof mats are 78.1,16, 17.5, 6.2, 22.9 and 26.1 respectively.

Table 1. Descriptive Statistics.

Index	consumption	Deliv-ery_points	height	floors	alt_prec	wall_mat	roof_mat
25%	29.1	11	12.7	5	1.5	10	10
50%	77.9	13	16.5	6	2.5	30	23
75%	115.6	18	21.9	7	2.5	30	40
count	55169	55169	54930	43552	55169	41922	41921
max	200	297	115.6	36	9999	95	94
mean	78.1	16	17.5	6.2	244	22.9	26.1
min	0.3	10	0	0	1	0	0
std	52.4	9.8	8.1	2.8	1536.2	17.6	23.7

3.3. Correlation Between Variables

To explore further associations among independent variables a correlation matrix is used. Table: 2 summarize this research findings. We can see a strong negative correlation between the sea surface height of the building and the level of gas consumption having $p = 0.01$. The other variables, delivery-points, height, floors, wall-mat and roof-mat have a negative effect on the dependent variable i.e., consumption.

DOI: 10.54254/2755-2721/2/20220648

Table 2. Correlation between Variables.

Index	Consumption	deliv-ery-points	Height	Floors	alt_prec	wall_mat	roof_mat
consump-tion	1.00	0.10	-0.18	-0.21	-0.01	-0.19	-0.24
deliv-ery_points	0.10	1.00	0.25	0.27	0.02	0.02	0.05
height	-0.18	0.25	1.00	0.82	-0.30	0.09	0.24
floors	-0.21	0.27	0.82	1.00	-0.08	0.12	0.28
alt_prec	-0.01	0.02	-0.30	-0.08	1.00	-0.04	-0.04
wall_mat	-0.19	0.02	0.09	0.12	-0.04	1.00	0.52
roof_mat	-0.24	0.05	0.24	0.28	-0.04	0.52	1.00

After applying feature engineering on our dataset, we found that consumption is the most important feature. The next important features are height and delivery points as shown in Fig. 2.

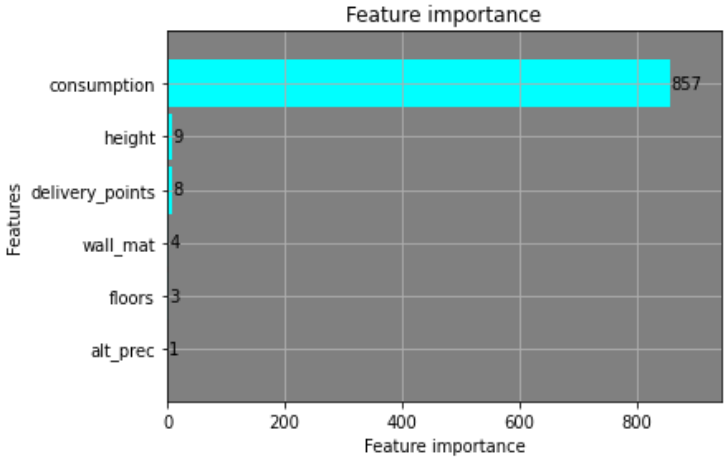

Figure 2. Features and their importance.

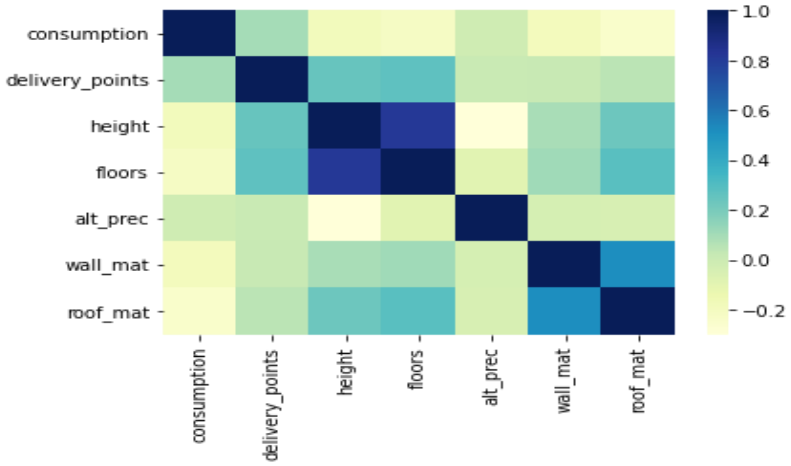

Figure 3. Heatmap of different features.

DOI: 10.54254/2755-2721/2/20220648

After finding out the correlation of different variables in the dataset, a heatmap is drawn to visualize the correlation. as seen in Fig.3.

Gas consumption is also largely affected by the type of building. Some of the different types of buildings in our dataset are Residential, Industrial, Commercial and Religious buildings. Fig. 4. shows the gas consumption level by building types. A graph has been plotted showing the consumption by the different types of buildings and it can be clearly seen in Fig 4: that the highest gas consumption has been done by the religious building types and followed by Industrial types. After applying feature engineering and utilizing the best features, we built our prediction model using a decision tree-based algorithm, Light GBM and Linear Regression. Light GBM uses a gradient boosting framework and is being used widely for building models. We used the evaluation criteria of Mean Square Error (MSE) value to evaluate proposed model. We obtained satisfactory results that could be utilized in our decision-making process to fight climate change and its adverse effects on the world.

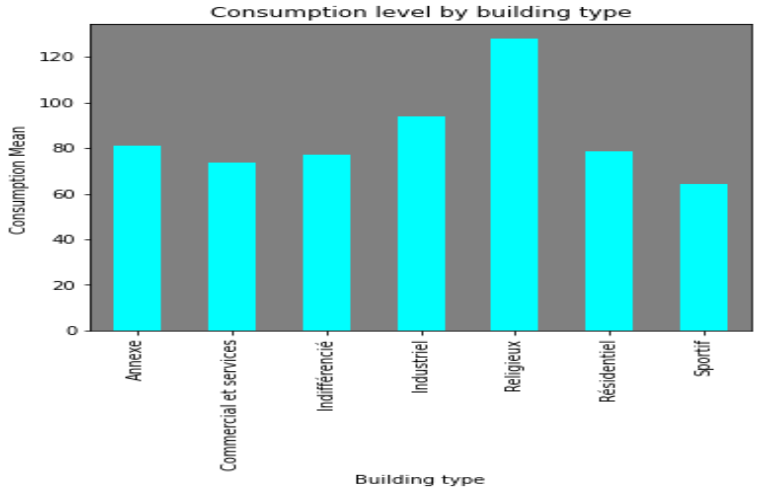

Figure 4. Consumption by each type of building.

Similarly, another study [13] that proposed a model to predict electrical energy consumption in buildings used a hybrid ARIMA-GBRT model and GBRT model, both being gradient boosting methods. They were also found to give best performance compared to the other techniques used. One of the research reported in [14] is also based on gradient boosting predicted energy consumption in commercial buildings. They utilized all the existing evaluation metrics to evaluate the model. Their results clearly showed that using the Gradient Boosting model enhanced the R-squared prediction accuracy and CV(RMSE) significantly compared to linear regression and random forest.

4. Conclusion

There is a huge demand for efficient, accurate techniques to forecast building energy usage to fight climate change. The conventional modelling techniques to predict the building energy consumption don't suffice the needs. As a matter of fact, machine learning algorithms have exhibited significant capabilities for predicting energy consumption for different types of buildings [15]. This paper focused on realizing the importance of various features, the correlation between them and their overall impact on our target variable i.e., consumption. This work used one of the most powerful algorithms owing to its high accuracy, Light GBM as the basis for the prediction model. As a result, it was found that the Light GBM algorithm achieved higher efficiency than the linear regression. Light GBM achieved a Mean Square Error value of 0.12 and RMSE value of 0.38. Hence, it can be utilized in our efforts to build decision systems to predict gas energy consumption in buildings.

As a future scope, this work can be extended thru integrating gradient boosting algorithms. The latest machine learning algorithms or their ensembles can also be used to build the prediction model. The

DOI: 10.54254/2755-2721/2/20220648

algorithms can then be further compared and the impact of various attributes on energy consumption in buildings can be analyzed.

References

[1] Shah, I., Akbar, S., Saba, T., Ali, S., & Rehman, A. (2021). Short-term forecasting for the electricity spot prices with extreme values treatment. IEEE Access, 9, 105451-105462.

[2] Khalid, A., Javaid, N., Mateen, A., Ilahi, M., Saba, T., & Rehman, A. (2019). Enhanced time-of-use electricity price rate using game theory. Electronics, 8(1), 48.

[3] European Environment Agency. Greenhouse Gas Emissions from Energy Use in Buildings in Europe 2021. Available online: https://www.eea.europa.eu/data-and-maps/indicators/greenhouse-gas-emissions-from-energy/assessment (accessed on 11 November 2021).

[4] Rehman, A., & Saba, T. (2011). Performance analysis of character segmentation approach for cursive script recognition on benchmark database. Digital Signal Processing, 21(3), 486-490.

[5] Rehman, A., & Saba, T. (2012). Off-line cursive script recognition: current advances, comparisons and remaining problems. Artificial Intelligence Review, 37(4), 261-288.

[6] Naz, A., Javed, M. U., Javaid, N., Saba, T., Alhussein, M., & Aurangzeb, K. (2019). Short-term electric load and price forecasting using enhanced extreme learning machine optimization in smart grids. Energies, 12(5), 866.

[7] Pathak, N., Ba, A., Ploennigs, J., & Roy, N. (2018, June). Forecasting gas usage for big buildings using generalized additive models and deep learning. In 2018 IEEE International Conference on Smart Computing (SMARTCOMP) (pp. 203-210). IEEE.

[8] de Keijzer, B., de Visser, P., Romillo, V. G., Muñoz, V. G., Boesten, D., Meezen, M., & Rahola, T. B. S. (2019). Forecasting residential gas consumption with machine learning algorithms on weather data. In E3S Web of Conferences (Vol. 111, p. 05019). EDP Sciences.

[9] Olu-Ajayi, R., Alaka, H., Sulaimon, I., Sunmola, F., & Ajayi, S. (2022). Building energy consumption prediction for residential buildings using deep learning and other machine learning techniques. Journal of Building Engineering, 45, 103406.

[10] Khan PW, Kim Y, Byun Y-C, Lee S-J. Influencing Factors Evaluation of Machine Learning-Based Energy Consumption Prediction. Energies. 2021; 14(21):7167.

[11] Panek W, Włodek T. Natural Gas Consumption Forecasting Based on the Variability of External Meteorological Factors Using Machine Learning Algorithms. Energies. 2022; 15(1):348.

[12] Yang, Y., Gong, N., Xie, K., & Liu, Q. (2022). Predicting Gasoline Vehicle Fuel Consumption in Energy and Environmental Impact Based on Machine Learning and Multidimensional Big Data. Energies, 15(5), 1602.

[13] Nie, P., Roccotelli, M., Fanti, M. P., Ming, Z., & Li, Z. (2021). Prediction of home energy consumption based on gradient boosting regression tree. Energy Reports, 7, 1246-1255.

[14] Touzani, S., Granderson, J., & Fernandes, S. (2018). Gradient boosting machine for modeling the energy consumption of commercial buildings. Energy and Buildings, 158, 1533-1543.

[15] Seyedzadeh, S., Rahimian, F. P., Glesk, I., & Roper, M. (2018). Machine learning for estimation of building energy consumption and performance: a review. Visualization in Engineering, 6(1), 1-20.