Modeling the energy consumption of buildings to support energy transition

ABSTRACT

This paper presents a model for predicting energy consumption in buildings in order to support the transition to more sustainable energy systems. Despite the significant impact of the real estate sector on greenhouse gas emissions and energy consumption, investments in energy efficiency and renewable energy for the housing sector remain insufficient. By providing real estate owners with accurate and actionable information on the potential energy savings from insulation and other energy efficiency measures, this model aims to enable and accelerate investments in energy transition. In this paper, we propose the implementation of an ML model to help predict the energy consumption of European buildings in order to estimate energy and cost savings, and automate energy assessments.

The results and tests show that the designed algorithm is correct and satisfies the criteria and performance.

Keywords: Machine Learning · Data Cleaning · Explained Variance · Feature Engineering

Overview

Background

Energy management is at the core of the sustainability of business, society, and our planet. Since the signing of the Kyoto Protocol in 1997 [1], the reduction of greenhouse gases and in particular CO2 emissions has become a major issue for all private and public actors. This is a key point if we hope to meet the objectives of the 2015 Paris Conference on Global Warming (COP21) [2] to limit global warming between 1.5°C and 2°C. On top of that, geopolitical issues have forced us to rethink our energy consumption to reduce our gas and electricity consumption.

According to a 2021 study, buildings alone account for 44% of the energy consumed in France, ahead of the transport sector (31.3%) while greenhouse gases emission account for 25% [3]. It is therefore crucial to understand which levers can be used to reduce the consumption of buildings and renovate them efficiently.

Project scope

Our project aims at leveraging the ML technology and the dataset of the European Buildings' yearly average energy consumption to develop a web-app to help building manager agents to monitor their long-term energy consumption and CO2 emission level in real time, diagnose the factors that they can improve to reduce both energy consumption and CO2 emission levels based on their refurbishment obstacles and priorities and propose accessible solutions based on their economic constraints and local law and regulations that can help them achieve the goal. In the end, our business plan includes building a local materials and labor market network/database to help our clients to most efficiently achieve the energy consumption and CO2 reduction goal.

Presentation of the group and task management

Our team includes four DSs that are focusing on ML product development: (Ben Jemaa Yosr, Derbel Yassine, Ben Ismail Rayen and Znaidi Dhia) and 2 product managers and business strategists (Dan Xie and Sonia Edouardoury). The two pillars of the team work closely together to interactively set and achieve our goals to a promising better building management for sustainable development.

METHOD

Data understanding

The data set is designed to help estimate energy and cost savings, and automate energy assessments for buildings in Europe. With 1.5 million instances of European buildings, it includes a wide range of technical features and yearly energy consumption data. By analyzing its features, we can gain a better understanding of how different factors impact energy consumption in buildings, which can help us make more accurate predictions about energy consumption and identify areas for improvement in terms of energy efficiency.

Data visualization is an essential tool for understanding and interpreting the relationships and correlations between features and labels in a data set. It allows the easy identification of patterns and trends and quickly identifies any outliers, anomalies or missing values. We chose to apply a log transformation to normalize features. It is a common technique used in machine learning and deep learning to improve the performance of models.

Data pre-processing

In the feature engineering process for this study, we dropped several features from the dataset because they were found to be highly correlated with other features. Specifically, we removed features nb_gas_meters_housing and nb_power_meters_housing due to correlation values of 1, and 1 respectively with other features. This decision was based on the fact that highly correlated features can lead to overfitting and can negatively impact the model's generalizability.

Additionally, we removed missing values from certain features that were deemed essential for the analysis. Specifically, we removed rows that contained missing values for some features. This was done using the dropna() function in pandas. The decision to remove rows with missing values in these features was based on the fact that they are essential in understanding the energy consumption of buildings, and missing values in these features would have led to bias in the analysis and inaccurate predictions.

We also performed a step to encode categorical variables present in the dataset. To do this, we first identified all the columns in the dataset that had a data type of "object", and stored them in a list called "categ". Then, we defined a function called "label_encoding", which takes in the dataset and the list of categorical variables as inputs. This step is important because many machine learning algorithms can only handle numerical values, so categorical variables must be transformed into numerical values to be used in the analysis.

We also performed a step to remove outliers from the dataset. Specifically, we removed rows from the dataset where the value of the feature 'energy_consumption _per_annum' was greater than 3200 as we couldn’t compare buildings with average energy consumption and others with high ones.

Model development and Deployment strategy

In this study, we will mainly focus on two different machine learning models, XGBoost and LGM, in order to achieve a good explained variance score. XGBoost, short for Extreme Gradient Boosting, is a popular model for regression and classification problems. It is an ensemble method that combines multiple decision trees to improve the performance of the model. LGM, short for Light Gradient Boosting Machine, is a lightweight version of XGBoost that is designed to be faster and more memory-efficient.

Carbon Footprint Limitation

This study does not consider the carbon footprint of the energy sources used to power the buildings. While the model proposed in this paper aims to support the transition to more sustainable energy systems by providing real estate owners with information on potential energy savings, it is important to consider the potential environmental impact of the real estate sector as a whole.

CONCLUSION

In summary, this study presents a machine learning model for predicting energy consumption in European buildings. The model is designed to support sustainable energy systems by providing information on potential energy savings. The study used XGBoost and LGM models and achieved a good explained variance score. However, the study does not take into account the broader environmental impact of the real estate sector and is limited to European buildings. Nevertheless, it provides valuable insights into the potential of machine learning in sustainable energy systems.

REFERENCES

[1] kyoto protocol signed: https://education.nationalgeographic.org/resource/kyoto-protocol-signed
[2]Paris agreement: https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement
[3] https://www.iea.org/topics/transport