Title: **Enhanced Building Energy Consumption Prediction Using a Dual-Model Approach and Ensemble Techniques**

Table of Contents:

1. Abstract 2

2. Introduction 3

3. Background and Motivation 3

4. Methodology 4

4.1 Data Collection 4

4.2 Data Extraction 5

4.3 Data Modelling 6

5. Results 8

6. Interpretations and Insights 9

Authorship:

This document was prepared by Rajashekar (Raja) Korutla, with significant contributions and guidance from Vamshi Gooje.

Abstract:

This study introduces a pioneering approach to predict building energy consumption, employing a dual-model framework based on Random Forest algorithms. Central to our methodology was the creation of a comprehensive dataset derived from HTML outputs generated by the EnergyPlus software. This process involved a rigorous data extraction phase, where both company-specific and open-source EnergyPlus outputs were meticulously processed. Through HTML scraping and PDF data extraction techniques, we constructed a rich dataset encompassing a wide array of energy-related parameters, laying the groundwork for a detailed analysis focused on factors influencing building energy consumption.

We divided our analysis into two distinct subsets: Subset A, emphasizing categorical variables related to HVAC systems and building types, and Subset B, focusing on a mix of continuous variables, building categories, and climate zones. This bifurcation allowed us to tailor our models to the unique characteristics of these variable types. Each subset was modeled using Random Forest Regressors, a choice driven by the method's robustness and ability to handle both categorical and continuous data efficiently. The models were rigorously evaluated using K-Fold cross-validation, ensuring the reliability and generalizability of our results.

Our findings reveal a notable difference in predictive performance between the two models. Subset B, with a mix of continuous and categorical variables, demonstrated a superior ability to predict energy consumption, as indicated by a high mean R^2 value of 0.9572 and a low mean MSE of 0.0377. In contrast, Subset A showed moderate predictive power with a mean R^2 of 0.6857.

To harness the strengths of both models, we implemented an ensemble approach, combining predictions from both models using a weighted average method, with greater emphasis on Subset B's predictions due to its higher accuracy. This ensemble technique not only capitalized on the individual models' strengths but also provided a more nuanced and comprehensive prediction of building energy consumption.

This study's methodology and findings hold significant implications for energy management in buildings, offering a robust model that can be utilized for energy efficiency analysis and decision-making. Our approach demonstrates the potential of machine learning techniques in enhancing the accuracy of energy consumption predictions, paving the way for more energy-efficient building designs and operations.

Introduction

In recent years, the need for efficient energy management in buildings has become increasingly crucial, driven by growing environmental concerns and the escalating costs associated with energy consumption. The analysis of building energy consumption using machine learning offers a promising avenue to address these challenges. Machine learning models, known for their ability to handle large and complex datasets, provide a sophisticated means of understanding and predicting energy usage patterns. This capability is particularly pertinent in the realm of building energy management, where optimizing energy usage can lead to significant economic and environmental benefits.

Background and Motivation

The traditional method of analyzing building energy consumption involves the use of simulation software like EnergyPlus. While EnergyPlus is robust and detailed, the time it takes to produce energy output simulations can be substantial, especially for large-scale studies. This limitation poses a challenge in scenarios where quick decision-making is crucial, such as in real-time energy management or in the early stages of building design, where multiple iterations are common.

Our project is motivated by the need to optimize the time and resources expended in these simulations. By employing machine learning models over the output data from EnergyPlus, we aim to develop a more time-efficient approach to predicting energy consumption in buildings. This method leverages the detailed data generated by EnergyPlus, while significantly reducing the time required to obtain actionable insights.

Objectives

The primary objective of our study is to predict various aspects of building energy consumption, such as heating, cooling, fan usage, pump operation, heat rejection, and the total energy consumed by lighting and equipment. To achieve this, we utilize a range of building-specific parameters extracted from EnergyPlus outputs. These parameters include:

- **Thermal Characteristics**: U-factors for walls and roofs, which measure the rate of heat transfer.
- **Building Physical Attributes**: Areas of walls, roofs, and the overall building area.
- **Building Design Elements**: Window-wall ratio, a critical factor in determining heat gain and loss.
- **Environmental Interactions**: Infiltration rates, which affect heating and cooling loads.
- **HVAC Equipment Types**: The presence of various heating, ventilation, and air conditioning systems.
- **Internal Loads**: Plug and lighting loads, representing the energy consumed by appliances and lighting fixtures.

- **Building Category**: The type of building, such as a hospital, office, or school, each having unique energy usage profiles.

By integrating these parameters into machine learning models, we aim to create a predictive framework that can quickly and accurately estimate energy consumption, thus facilitating more effective energy management strategies in buildings.

Methodology

Data Collection

The data collection phase of the project involved acquiring a substantial set of HTML output files from EnergyPlus simulations, crucial for the analysis of building energy consumption. This phase can be detailed in two main parts:

1. **Company-Specific Data Collection:**
    - **Source Description**: The initial set of data was collected from a variety of projects handled by Thornton Tomasetti.
    - **Data Characteristics**: This collection comprised approximately 600 records, predominantly from projects located in climate zones 5a and 5b, with a significant focus on school-based projects.
    - **Collection Process**: The data was gathered through coordination with engineers who worked on these projects. This process involved obtaining both simulation files and real project files.
    - **Variation in Data**: Notably, the dataset included files where few parameters were altered in the same project to simulate different scenarios, adding diversity to the data.
2. **Open-Source Data Collection:**
    - **Source Description**: In addition to the company-specific data, an open-source collection of HTML outputs from EnergyPlus was utilized.
    - **Data Scope**: This open-source dataset included prototype models for 16 different commercial building types across 19 climate locations, which encompassed both U.S. and international locations.
    - **Standards and Versions**: The models conform to recent editions of ASHRAE Standard 90.1 and IECC, employing EnergyPlus™ Version 22.1.0.
    - **Dataset Size**: The open-source collection contributed to an expansive dataset with 3,952 total building models.

By combining data from both company-specific projects and an extensive open-source dataset, this study leverages a rich and diverse array of building energy simulations. This comprehensive data collection strategy ensures a robust foundation for the subsequent analysis and modeling phases of the study.

Data Extraction And Modelling

The data extraction phase of the project was critical in shaping the dataset that formed the basis of your analysis. This phase can be expanded to include specific details about the process, particularly focusing on the Python script and its functionality:

1. **Collaboration with Experts**:
    - **Interdisciplinary Input**: Regular interactions with the engineering team, who possess expertise in energy modeling, were key. These discussions helped in understanding the nuances of EnergyPlus outputs and in identifying the most impactful parameters for energy consumption analysis, such as U factors, building areas, and HVAC types.
    - **Parameter Selection**: The collaborative effort led to the identification of parameters that have a direct influence on building energy consumption. This step was crucial in ensuring that the data extracted was relevant and comprehensive.
2. **Development of Data Extraction Program**:
    - **Script Development**: A custom Python script was developed to extract the desired features from the HTML reports. The script was designed to search for specific keywords associated with the identified parameters.
    - **Iterative Refinement**: The script underwent several rounds of refinement, adjusting the parameters being extracted or transformed based on continuous feedback and new insights.
3. **Automated Extraction and Structuring of Data**:
    - **Extraction Using Python Libraries**: The script employed BeautifulSoup for HTML parsing, which enabled the extraction of diverse data points from the HTML files, including climate zones, building types, and various energy-related metrics. For PDF data, pdfplumber was used to extract relevant information.
    - **Data Aggregation and Structuring**: All the extracted data was compiled into a structured Pandas DataFrame. This step was crucial in transforming the varied raw data into a consistent and analyzable format.
4. **Dynamic Data Handling**:

- **Adapting to Data Variability**: The script was designed to be flexible enough to handle different data structures and formats due to the diversity of the source files. This flexibility was necessary to accommodate the variations in HTML structures across different EnergyPlus versions and project files.
- **Continuous Script Improvement**: The extraction process was dynamic, requiring ongoing modifications to the script as new requirements and insights emerged. This aspect of the project highlights the adaptive nature of the data extraction process.

5. **Specific Script Functionalities**:

   Climate Zone Extraction: The script identified climate zones by extracting the WMO number from HTML files and correlating it with the appropriate climate zone.

6. **Building and Energy Feature Extraction:**

   - **U-Factors:** Extracted U-factors for walls and roofs from specific HTML tables.
   - **Building Areas:** Determined net wall and roof areas from the data.
   - **HVAC Types:** Identified and extracted various HVAC equipment types.
   - **Complex Data Relationships:** Managed relationships between building types, energy consumption patterns, and HVAC systems.
   - **Geothermal Loop Presence**: Checked for "GEOTHERMAL LOOP" or "HEAT PUMP LOOP" in files.
   - **Building Type Determination:** Classified building types, such as distinguishing schools by identifying specific keywords like "GYM_"**.**
   - **Energy Metrics Extraction:** Extracted detailed metrics like glass U-factor, SHGC, infiltration rates, window-wall ratios, plug and process loads, lighting loads, and more.
   - **HVAC Equipment Analysis:** Analyzed the presence of specific HVAC equipment types based on energy consumption patterns.

7. **Data Aggregation and Transformation:**

- **Comprehensive Data List:** The script compiled a data list, appending each file's extracted information, including building type, climate zone, and various energy metrics.
- **DataFrame Creation:** A Pandas DataFrame was created from the aggregated data list, ensuring a structured and accessible format for further analysis.

This detailed explanation of the data extraction phase underscores the technical proficiency in programming and data handling. The use of Python and its libraries facilitated a comprehensive and efficient extraction process, while the continuous collaboration with domain experts ensured the relevance and accuracy of the extracted data. This phase laid a strong foundation for the subsequent analysis and was pivotal in ensuring the success of the project.

Check out the python script here.

8. **Data Preparation and Feature Selection:**
    - we started by loading the dataset and dropping unnamed and redundant columns, as part of data cleaning.
    - Features for two subsets, A and B, were defined based on different types of variables (HVAC types and building categories for Subset A; continuous variables, building categories, and climate zones for Subset B).
    - The target variables, which include various energy consumption metrics, were clearly identified.
9. **Model Initialization and Training:**
    - we initialized two Random Forest models, one for each subset.
    - Before training, we applied a log transformation (log1p) to the target variables, which is commonly used to handle skewness in the data.
    - Cross-validation using K-Fold (with 5 splits) was performed to assess the model performance, measuring both R^2 and MSE scores. This step is crucial for evaluating the model's generalizability and robustness.
10. **Feature Importance Analysis:**
    - Post-training, we extracted feature importances from both models. This is a valuable step for understanding which features have the most influence on the predictions.
    - The feature importances were sorted and displayed, offering insights into the relative significance of different features in the model.
11. **Prediction and Integration:**
    - we used the trained models to predict new data.
    - The predictions from both models were combined using a weighted average approach. The weights (0.30 for Subset A and 0.70 for Subset B) suggest a higher reliance on the predictions from Subset B.
12. **Final Output:**
    - Finally, the combined predictions were displayed, providing the final output of your modeling process.

[Access the machine learning model here]

13. **Results:**

Random Forest on Subset A:

- **Mean R^2**: 0.6857, indicating a moderate proportion of variance in the target variable is predictable from the input variables.
- **Std R^2**: 0.0115, showing relatively low variability in the R^2 metric across different folds.
- **Mean MSE**: 0.2955, representing the average squared difference between observed and predicted values.
- **Std MSE**: 0.0119, indicating the model's consistency in terms of error.

Random Forest on Subset B:

- **Mean R^2:** 0.9572, suggesting a high proportion of variance in the target variable is predictable, indicating a better fit than Subset A.
- **Std R^2**: 0.0042, showing very low variability in R^2 across folds.
- **Mean MSE**: 0.0377, a significantly lower error metric compared to Subset A.
- **Std MSE**: 0.0023, demonstrating high consistency in the model's performance.

Feature Importance Analysis:

Subset A:

	Feature	Importance
7	Building_Type_Restaurant	0.396411
4	Building_Type_Hospital	0.189929
1	Boiler	0.113357
10	Building_Type_Warehouse	0.076943
8	Building_Type_Retail	0.058624
6	Building_Type_Office	0.044181
9	Building_Type_School	0.038140
2	DX_Cooling	0.033144
3	DX_Heating	0.022960
0	Chiller	0.020116
5	Building_Type_Lodging	0.006197

The most influential feature in Subset A was 'Building Type Restaurant', followed by 'Building Type Hospital' and 'Boiler'. This indicates that building type plays a significant role in the model's predictions.

Subset B:

	Feature	Importance
10	Building_Area__ft2_	0.270694
5	Glass_SHGC	0.230073
8	Plug_and_Process__W_ft2_	0.137896
6	Infiltration__cfm_ft2_	0.075174
9	Lighting__W_ft2_	0.053870
11	vpz__cfm_	0.050259
7	Window_Wall_Ratio	0.035187
13	Climate_Category_cool	0.027260
12	vot__cfm_	0.021808
15	Climate_Category_temperate	0.020726
2	Net_Wall_Area__ft2_	0.020085
4	Glass_U_Factor__Btu_hr_ft2__F_	0.019254
0	U_Factor_Wall__Btu_hr_ft2__F_	0.018642
3	Net_Roof_Area__ft2_	0.009880
1	U_Factor_Roof__Btu_hr_ft2__F_	0.007751
14	Climate_Category_hot	0.001442

In Subset B, 'Building Area [ft²]' and 'Glass SHGC' emerged as the top features, highlighting their importance in predicting building energy consumption.

Interpretation and Insights

The model's performance and feature importance analysis provide valuable insights into factors influencing building energy consumption. The significant role of building type in Subset A suggests variations in energy usage patterns across different building categories. In Subset B, the building's physical characteristics, such as area and glass properties, are more predictive of energy consumption, indicating their crucial role in energy modeling.